



# The Connectionist Learning Of Mental Representation

**MICHAEL DAVID LEE**

B.Sc. (Ma.), B.A. (Hons.)

Thesis submitted for the degree of

**Doctor of Philosophy**

**The University of Adelaide**  
Departments of Psychology and  
Electrical and Electronic Engineering

May 1997

# TABLE OF CONTENTS

<b>Chapter 1: Mental Representation And Connectionist Modelling</b>	<b>1</b>
1.1. The Connectionist Modelling Framework	1
1.1.1. Emergent Cognitive Phenomena	2
1.1.2. Embodied And Situated Cognition	3
<b>Chapter 2: Mental Representation In Connectionist Models</b>	<b>6</b>
2.1. Representation Through Pre-Abstracted Features	7
2.1.1. Connectionist Models Employing Featural Stimulus Representations	7
2.1.2. The Problems With Pre-Abstracted Psychological Features	9
2.2. Representation Through Sensory Description	10
2.2.1. Connectionist Models Employing Sensory Stimulus Representation	10
2.2.2. Evaluating The Sensory Description Representational Approach	12
2.3. Connectionist Models Which Learn Internal Representations	14
2.3.1. Connectionist Semantic Networks	15
2.3.2. The Semantic Map	18
2.3.3. Evaluating The Connectionist Models	19
2.3.4. Conclusion	21
<b>Chapter 3: Psychological Space</b>	<b>22</b>
3.1. Foundations Of Psychological Space	22
3.1.1. The Construction Of Psychological Spaces By Multidimensional Scaling	23
3.1.2. The Universal Law Of Generalization	25
3.1.3. Distance Metrics In Psychological Space	27
3.1.4. Empirical Evaluations Of Psychological Space Representations	29
3.1.5. Criticisms Of The Psychological Space Position	30
3.2. Connectionist Models Using Psychological Space Representations	35
3.2.1. The Consequential Region Model	35
3.2.2. Shepard And Kannappan's Model	36
3.2.3. The Radial Basis Function Approach	38
3.2.4. The ALEX Model	39
3.2.5. The ALCOVE Model	40
3.2.6. Comparing The Consequential Region And Radial Basis Function Approaches	41
3.3. The Connectionist Learning Of Psychological Space Representations	43
3.3.1. Rumelhart And Todd's Model	44
3.3.2. Connectionist Multidimensional Scaling In A Radial Basis Function Architecture	46
<b>Chapter 4: A Connectionist Multidimensional Scaling Model</b>	<b>48</b>
4.1. Processing Phase	49
4.1.1. Stimulus Presentation	49
4.1.2. Determining The Current Internal Representation	50
4.1.3. Calculating Psychological Similarity	51
4.1.4. Provision Of Feedback	53
4.2. Learning Phase	54
4.2.1. Similarity Error	54
4.2.2. Dimensional Error	54
4.2.3. Derivation Of The Learning Rule	59
4.3. Construction And Interpretation Of The Model	62
<b>Chapter 5: Evaluation Of The Connectionist Multidimensional Scaling Model</b>	<b>65</b>
5.1. Demonstrations Of The Model	65
5.1.1. Colour Model	65
5.1.2. Flower Pot Model	67
5.2. Evaluation Of The Model	69
5.2.1. Sensitivity To Parameter Values	70
5.2.2. Overcoming Separable Stimulus Difficulties	74
5.2.3. Entrapment In Local Minima	77

<b>Chapter 6: The Internal Derivation Of Psychological Similarity</b>	<b>83</b>
<b>6.1. The Relationship Between The Mind And The World</b>	<b>83</b>
6.1.1. The Rational Approach	86
6.1.2. Psychological Essentialism	88
6.1.3. Psychophysical Complementarity	90
6.1.4. Process-Based Representation	91
6.1.5. Physical Constraints	93
6.1.6. Conclusion	95
<b>6.2. Connectionist Internalisation Of Psychological Similarity</b>	<b>96</b>
6.2.1. Categorical Associations And Sensory Properties	96
6.2.2. A First Approximation Of Psychological Similarity	100
6.2.3. An Anticipatory Rejoinder	101
<b>Chapter 7: A Model Of The Learning Of Mental Representation</b>	<b>104</b>
<b>7.1. Processing Phase</b>	<b>105</b>
7.1.1. Response Generation	105
7.1.2. Target Similarity Derivation	105
7.1.3. Environmental Feedback Provision	108
<b>7.2. Learning Phase</b>	<b>109</b>
7.2.1. External Error	109
7.2.2. Internal Error	110
<b>7.3. Model Construction And Parameter Setting</b>	<b>111</b>
7.3.1. Basis For Setting The Information Parameter	112
7.3.2. Distribution Of Target Similarities	113
7.3.3. Maximising Entropy	116
<b>7.4. Overview Of The Mental Representation Learning Model</b>	<b>121</b>
7.4.1. Relationship To Connectionist Semantic Networks And The ALEX Model	121
7.4.2. Relationship To Piagetian Learning Principles	124
<b>Chapter 8: Demonstrations Of The Mental Representation Learning Model</b>	<b>127</b>
<b>8.1. Sensory Properties</b>	<b>127</b>
8.1.1. The Bug Model	127
8.1.2. The Berry Model	130
<b>8.2. Categorical Associations</b>	<b>133</b>
8.2.1. The Bird Model	134
8.2.2. The Animal Model	136
8.2.3. The Senator Model	139
<b>8.3. Adaptation To A Dynamic Environment</b>	<b>143</b>
8.3.1. The Dynamic Bug Model	143
8.3.2. Adaptational Possibilities And Interpretations	145
<b>Chapter 9: Extensions To The Mental Representation Learning Model</b>	<b>148</b>
<b>9.1. Model Refinements</b>	<b>148</b>
9.1.1. Adaptive Parameter Setting	148
9.1.2. Selective Attention Response Space Weighting	149
9.1.3. Learning The Metric Structure Of Psychological Space	150
9.1.4. Response Space Basis Function	152
9.1.5. Dimensional Error	155
<b>9.2. Learning Psychophysical Mappings</b>	<b>159</b>
9.2.1. Multidimensional Scaling Model	160
9.2.2. Mental Representation Learning Model	162
<b>Chapter 10: Concluding Remarks</b>	<b>165</b>
<b>REFERENCE LIST</b>	<b>168</b>

The connectionist modelling framework enables the construction of cognitive models which are situated, embodied, and support emergent cognitive phenomena. Coupled with the potential to learn complicated internal representations, these abilities endow connectionism with significant promise as a means of modelling the nature, acquisition and utilisation of mental representations. After arguing for the validity of the established 'psychological space' theory of mental representation, this thesis develops and evaluates two models which explore the connectionist learning of this type of representation. The first model provides the groundwork for the second, by demonstrating a way in which the representational structures dictated by the psychological space theory may be learned by a connectionist network. This model effectively implements a metric multidimensional scaling algorithm by assuming that psychological similarity exponentially decays in relation to distance in psychological space. The model is shown to be capable of operating under the Minkowskian family of distance metrics, and autonomously determines the appropriate dimensionality of the derived representational space. The second model attempts to provide a more realistic account of the learning of psychological spaces by internally generating the similarity indices provided externally to the first model. It is argued that such indices may be derived from representational constraints implicit in the cognitive operation of the model in its environment, particularly from information regarding the categorical associations and sensory properties of stimuli. This second model is shown to be capable of deriving appropriate representations through learning these associations and properties, and its behaviour is also demonstrated in noisy and dynamically changing environments. The relationship of the models both to general psychological learning theory and to other connectionist models is discussed, and several possible extensions and refinements to the models are explored.

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

This thesis was written between February 1994 and April 1997, primarily within the Psychology Department at the University of Adelaide under the supervision of Associate Professor Douglas Vickers. In February 1996, however, I joined the postgraduate program of the Cooperative Research Centre for Sensory Signal and Information Processing (CSSIP), and acquired Professor Robert Bogner from the Electrical and Electronic Engineering Department of the University of Adelaide as a co-supervisor. I would like to thank both Doug and Bob for their interest and assistance in relation to this thesis, and for the more general career opportunities they have provided. Besides my supervisors, various discussions with a number of people have contributed to the substance of this work: I would particularly like to thank Kenneth Pope, Charles Pearce, Philip Smith, Jon Baxter, and the participants in the CSSIP research 'gatherings'.

I would also like to take this opportunity to thank my parents, David and Colleen, for their lifelong support, encouragement and guidance. Finally, and most especially, I would like to thank Elisa, my fiancée, for her unfaltering devotion, companionship and love.

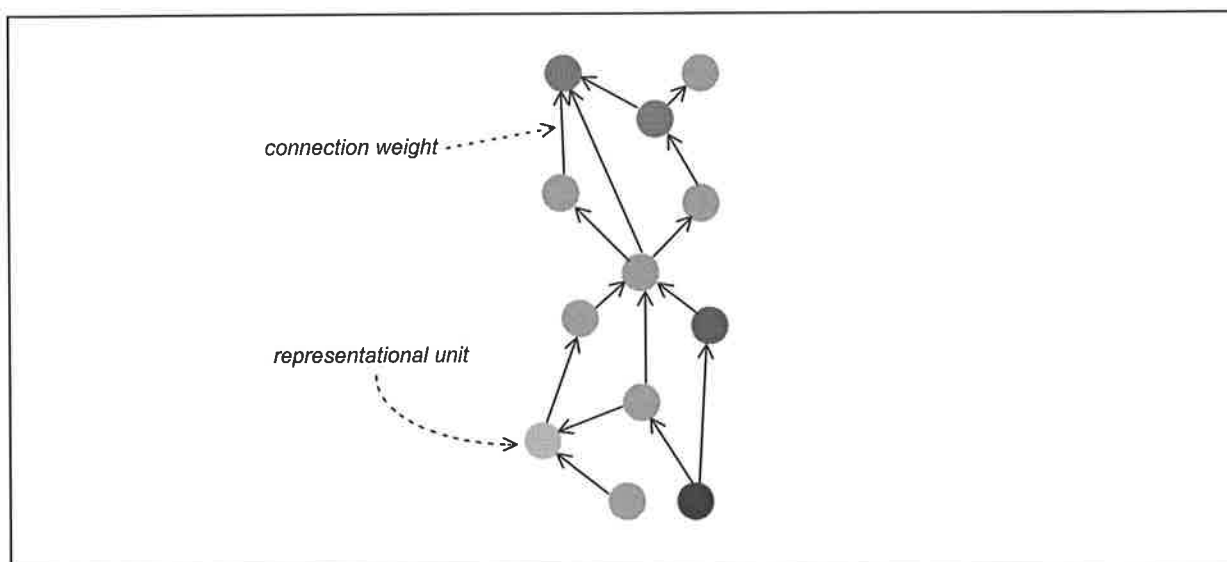
Michael Lee, Adelaide, South Australia, May 1997

# Chapter 1: Mental Representation And Connectionist Modelling

There is nothing more important to human mental life than its representational foundations. Mental representations provide the information structures with which we characterise and comprehend the external world, through which we plan and preview our interaction with that world, and upon which we can create and construct imaginary objects, events, and other worlds. So pervasive is the role of mental representation that an understanding of what it is, how it is learned, and how it is used is necessary - it has almost been suggested sufficient (Hofstadter 1988) - for the explanation, prediction, and emulation of human cognition.

This thesis examines the nature, acquisition and utilisation of mental representation through describing and evaluating two connectionist models. The first model provides the groundwork for the second by developing a way in which mental representational structures may be accommodated within a connectionist network. As such, the first model primarily considers the *nature* of mental representation. The second model, however, also incorporates the inter-related abilities of *acquisition* and *utilisation*, and thus constitutes a more complete attempt to model the learning of human mental representation.

## 1.1. The Connectionist Modelling Framework



**Figure 1.1.** A connectionist network. Different shadings of the representational units correspond to different activation values.

Since all of the modelling undertaken in this thesis is conducted within a connectionist framework, it is worthwhile providing a brief summary of this framework, and examining its cognitive modelling advantages. As shown in Figure 1.1, connectionist models consist of simple representational elements, referred to as 'units' or 'nodes', which are linked by a set of directed 'connection weights' to form a network. Information is represented within the network by associating a numerical 'activation value' with each unit, which is determined both by the

activation values of connected units, and the strengths of these connections. Connectionist models are also subject to learning rules which act to modify connection weights, thus altering the information processing properties of the network.

The ‘neural inspiration’ underlying the connectionist modelling framework - in which units correspond to generic neurons and connection strengths model synaptic junctions (see Rumelhart 1989, pp. 133-136) - is entirely appropriate, given the assumed biological basis of mental representation. This is not to suggest, however, that the models presented in this thesis are intended to be biologically realistic, or even biologically plausible. Some connectionist research (eg. Lynch, Granger, Larson & Baudry 1989, McClelland, O'Reilly & McNaughton 1995) strives for the detailed alignment of connectionist models with neurological brain structures. In this thesis, however, connectionism is employed merely as a modelling vehicle with which to develop a *psychologically* based understanding of the nature, acquisition and utilisation of mental representations. In this sense, the approach adopted here accords with Hofstadter's (1985) belief that:

“A model of thought ... need not be based so literally on brain hardware that there are neuron-like units and axon-like connections between them” (p. 659)

Rather, the two models developed in this thesis are intended to exist at what Smolensky (1988a, 1988b) terms the ‘sub-symbolic’ level, which is asserted to be distinguishable from a ‘neural’ modelling level. From this perspective, the fundamental advantages of connectionist modelling lie in its ability to model emergent cognitive phenomena, and its accommodation of embodied and situated cognition.

### 1.1.1. Emergent Cognitive Phenomena

Through their compatibility with the sub-symbolic approach, connectionist models are naturally able to accommodate the notion that mental representations are active and emergent phenomena (see Hofstadter 1985, chap. 26). This view derives from observed deficiencies in the traditional symbolic approach to cognitive modelling (eg. Newell 1980, Newell & Simon 1976). As summarised by Chalmers, French and Hofstadter (1991), the symbolic approach:

“posits that thinking occurs through the manipulation of symbolic representations, which are composed of atomic symbolic primitives” (p. 6)

corresponding to mental representations. The weakness of this conceptualisation is that the mental representational symbols are modelled as passive and inflexible data structures, reliant upon the action of external processing to capture meaning.

The alternative sub-symbolic approach to modelling human cognition argues that:

“cognitive behavior emerges as a statistical property of many small things designed to interact with one another” (Casti 1989, p. 305)



In other words, sub-symbolic models do not attempt to explicitly model either cognitive processes or mental representations, but seek instead to provide a formal computational substrate from which these features of cognition emerge as collective phenomena. Importantly, this substrate can be constructed from a set of *local* interactions between units, meaning that the need for an overarching cognitive executive is alleviated. Similarly, within the sub-symbolic approach, mental representations are conceived of as:

“ACTIVE SUBSYSTEMS of a complex system, and they are composed of lower-level active subsystems ... They are therefore quite different from PASSIVE symbols, external to the system ... which sit there immobile, waiting for an active system to process them” (Hofstadter 1979, p. 326)

As detailed by Smolensky (1988a, 1988b) and Cussins (1990) amongst others, connectionist modelling is entirely compatible with the sub-symbolic approach. Whilst the representational capabilities of units in a connectionist network are too limited to maintain symbolic representations, an interconnected concert of units is well suited to realising emergent representational structures. Thus, within connectionist models, it is patterns of activation across sets of units, rather than particular activation values of individual units, which are amenable to psychological interpretation as mental representations. As Smolensky (1988a) summarises:

“cognitive descriptions [are] built up of entities that correspond to constituents of the symbols used in the symbolic paradigm; these fine-grained constituents ... are the activities of individual processing units” (p. 3)

Accordingly, the mental representations learned by both of the models developed in this thesis are distributed across a number of units. The adherence to the principle of emergent cognitive modelling this promotes is evident in at least two tangible benefits. First, the distribution of representational information enhances the ability of the models to generalise (see, for example, French 1991, Hinton, McClelland & Rumelhart 1986, Kruschke 1993b). Secondly, the fact that a number of units are involved in any given mental representation assists in relating the models to the geometric ‘psychological space’ view of mental representation (see Shepard 1987a) which is espoused in this thesis.

### 1.1.2. Embodied And Situated Cognition

Connectionism can also readily accommodate a ‘situated’ approach to cognitive modelling (see Brooks 1991a, 1991b, Norman 1993, Russell & Norvig 1995, chap. 2) which asserts that cognition can only exist within an embodied agent operating in a structured external environment. This belief has its origins in observing that mental representations continually interact with the external world in that they form as adaptations to the external world, and in that the actions they induce directly affect the external world. As Norman (1993) notes, the situated approach to cognition focuses:

“entirely upon the structures of the world and how they constrain and guide human behavior. Human knowledge and interaction cannot be divorced from the world” (p. 4)

Connectionist models, through specifying subsets of ‘input’ and ‘output’ units within the network, are capable of being both embodied and situated. Input or ‘sensor’ units may be likened to human perceptual systems in the sense that they receive information from an external source, whilst output or ‘effector’ units may be likened to human motor systems in that they are able to alter that information source. The remaining units are generally termed ‘hidden’ units, and mediate the flow of information between the input and output units.

In the context of modelling situated cognition, a real or simulated external environment is appropriately employed as the external information source. In this way, as indicated in Figure 1.2, the external world may interact with, and be affected by, the connectionist model. As is also shown in Figure 1.2, it is natural to endow embodiment upon such models, through identifying the points at which exchanges of information take place with the boundary between the model and the external world.

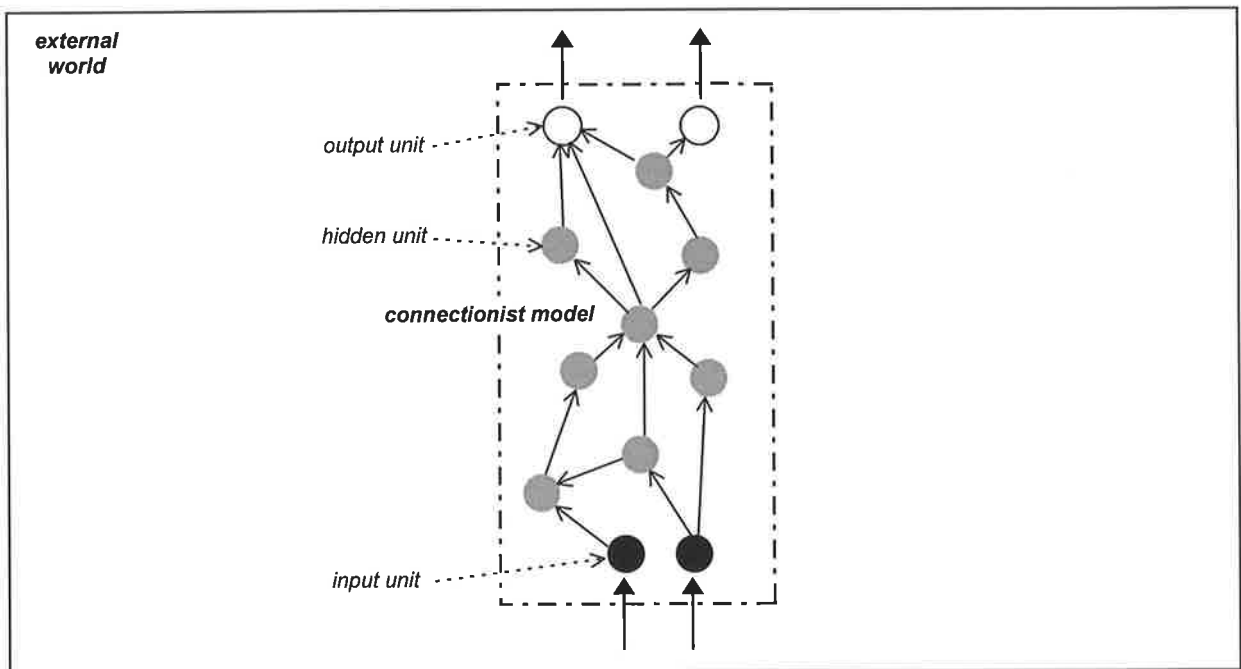


Figure 1.2. An embodied and situated connectionist model, incorporating input, hidden, and output units.

The second model developed in this thesis, which considers the acquisition and utilisation of mental representations, is heavily reliant upon information provided by an external environment. As such, the model is constructed as an embodied and situated cognitive agent, operating in a structured world, and incorporates input, output and hidden units fulfilling precisely the roles described by Figure 1.2.

On the basis of these considerations, therefore, the connectionist framework appears to be well suited to the modelling of the learning of human mental representation. Not only does connectionism provide a means for the interaction between a cognitive agent and the world to be

explicitly modelled, but the distribution of information throughout a connectionist network offers the promise of enabling mental representations to emerge naturally as an result of this interaction.

## Chapter 2: Mental Representation In Connectionist Models

This chapter provides a summary of the ways in which mental representations have been incorporated into previous connectionist modelling. It has convincingly been argued (eg. Hanson & Burr 1990, Hinton 1989) that much of the impetus for the widespread adoption of the connectionist modelling framework can be seen as the result of the development of learning procedures which allow connectionist models to develop appropriate internal representations. As such, part of this chapter is concerned with examining those networks which attempt to model the development of mental representations by coupling these learning procedures with particular network architectures.

The connectionist framework has, however, also been more generally employed in the modelling of psychological processes such as identification (eg. McClelland & Rumelhart 1981), categorisation (eg. Shanks 1991), and selective attention (eg. Kruschke 1992). Although the primary focus of these models is, naturally, on the process or processes which they are attempting to model, they do, necessarily, involve the specification of some form of mental representational structure. A complete examination of mental representation in connectionist modelling must, therefore, also consider the representational approaches of these models.

With regard to this second source of connectionist representational approaches, a legitimate initial reservation might be that the focus on different psychological processes would serve to confound the representational structures extracted from the models. Fortunately, many of the models under consideration are appropriately regarded as members of the broad class of 'cognitive process models' (Nosofsky 1992). In essence, cognitive process models are ones in which cognitive processes operate on fixed stimulus representations. Cognitive process modelling, for example, readily allows the same representational structure to be employed in both identification and categorisation tasks, with different transformational processes being employed to produce the required performance for each task. Similarly, categorisation data across two different sets of stimuli can be modelled by altering the representational structure, but leaving unchanged the categorisation processes which act upon these stimulus representations. Thus, although both a transformational process and a representational structure are necessary to model identification, categorisation, or any other type of cognitive performance, these two components of a cognitive process model maintain a certain degree of independence. As such, it is possible to survey the representational structures employed in previous connectionist modelling of this type, without being overly susceptible to the danger of the different aims of these previous models confounding the survey's conclusions. Accordingly, this chapter commences by examining the two dominant approaches to incorporating mental representation into connectionist cognitive process models.

---

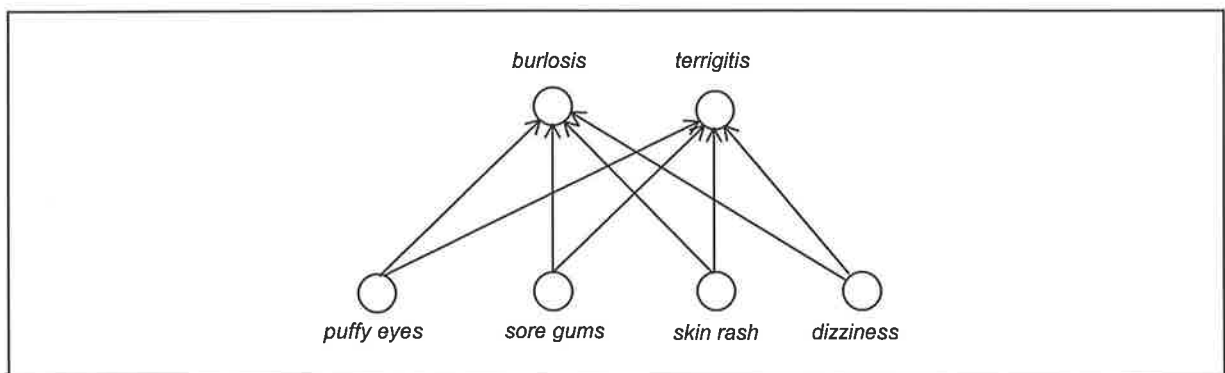
## 2.1. Representation Through Pre-Abstracted Features

Perhaps the simplest approach to incorporating mental representation into a connectionist model is to measure each stimulus in terms of a set of relevant psychological features. For example, a set of stimuli drawn from the natural kind ‘animals’ can be characterised in terms of their size, how many legs they typically have, their colouring, whether they have webbed feet, or gills, or claws, and so on. Within the connectionist model, allocating each of these features to a sensor (or input) representational unit provides a simple means of indicating to a model that it is encountering a certain animal in the environment: these featural units need only be set to appropriate, pre-coded, values in accordance with the psychological features of each animal.

### 2.1.1. Connectionist Models Employing Featural Stimulus Representations

This form of connectionist representational structure, despite its obvious shortcomings, has been widely employed in connectionist models. Three such models are described below to clarify the nature of this approach, and to allow a discussion of its weaknesses. Furthermore, the learning algorithms and modelling goals of these three models vary substantially, giving some indication of the pervasive breadth of this representational practice.

Shanks (1991) describes a series of connectionist models which learn to diagnose each of a set of patients into one of a number of disease categories on the basis of each patient’s physical condition. This condition is described in terms of the presence or absence of a finite number of symptoms, such as ‘skin rash’, ‘puffy eyes’, ‘dizziness’, and so on. The architecture of these models is particularly simple, and is shown in Figure 2.1.

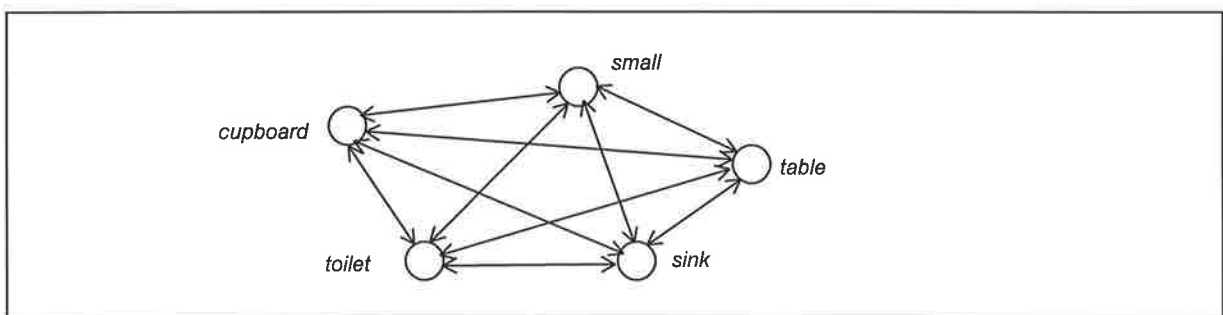


*Figure 2.1. The general architecture of Shanks' (1991) disease diagnosis models. Adapted from Shanks (1991), Figure 1.*

Each unit in the bottom ‘layer’ of these models corresponds to one of the symptoms, and, for each patient the model encounters, their physical condition is represented by the pattern of activation values across this layer. The presence of a particular symptom results in the appropriate unit being activated, whilst its absence leaves the unit inactive. Clearly, the representations utilised by the model in learning to diagnose the patients are solely based upon a set of pre-abstracted psychological features, in the form of symptoms.

A second example is the ARTMAP model (Carpenter, Grossberg & Reynolds 1991), based on Adaptive Resonance Theory (for an overview, see Carpenter 1989, Carpenter & Grossberg 1988), which learns to categorise a set of mushroom stimuli as either poisonous or edible. The architecture of the ARTMAP model is considerably more complicated than that of Shanks' (1991) model, and the learning methods employed are also entirely different, but the representational approach is virtually identical. Each mushroom is presented to the model as a bit string, with each bit encoding the presence or absence of one of 126 different mushroom features such as 'conical cap shape', 'green cap colour', 'fishy odour', 'urban habitat', and so on. Again, the representational structure of the model is one based entirely upon the presence or absence of a set of pre-abstracted psychological features.

A final example is Rumelhart, Smolensky, McClelland and Hinton's (1986) demonstration of a model of human conceptual structure, formalising a connectionist interpretation of previously established psychological constructs such as schemata (Rumelhart 1980), scripts (Schank & Abelson 1977), and frames (Minsky 1975, 1986). The model realises the emergence of room schemata, such as 'kitchen', 'dining room', and 'bathroom', in terms of household furniture and room properties such as 'small', 'cupboard', and 'table'. The architecture of this model is shown in Figure 2.2.



*Figure 2.2. The architecture of Rumelhart, Smolensky, McClelland, and Hinton's (1986) room schemata model. Only five of the forty, completely interconnected, featural room descriptors are shown.*

Each unit in Figure 2.2 corresponds to a piece of furniture or a room property, a number of which are initially activated, and are 'clamped' to remain active, indicating that they are part of the current room description. The model is then iteratively updated, transferring activation values between units through the connection weights, until the activation value at each unit stabilises as either entirely active or entirely inactive. The connection weights between each pair of units are set according to a Bayesian analysis (see Hinton & Sejnowski 1983) of the probability that the two pieces of furniture or room properties represented by the units co-occur in any given household room. The stable activation state of the model represents the model's generation of a completed room description. Rumelhart et. al. (1986) convincingly argue that these room descriptions can also be viewed, at a psychological level, as the outcomes of a schematic memory structure. They view the model's information processing technique, which is essentially a constraint satisfaction algorithm, as allowing:

“schemata [to] emerge at the moment they are needed from the interaction of large number of much simpler elements all working in concert with one another” (Rumelhart et. al. 1986, p. 20).

In terms of the schematic room concepts, ‘kitchen’, ‘dining room’, ‘bathroom’, and so on, this model is clearly employing a more sophisticated representational approach than either Shanks’ (1991) model, or the ARTMAP model. Indeed, the generation of emergent mental representations in this way was nominated as one of the principal attractions of the connectionist modelling approach in Chapter 1. As has been noted by Dyer (1988) and McCarthy (1988), however, the sub-conceptual room ‘microfeatures’ such as ‘small’, ‘cupboard’, and ‘table’ are represented by the model in a fundamentally different manner. In fact, they take the form of pre-abstracted psychological features.

### 2.1.2. The Problems With Pre-Abstracted Psychological Features

Whilst the practice of realising mental representation in connectionist models through identifying units with psychological features has the advantage of being straight-forward and transparent, it is, in many ways, an unsatisfactory approach. The representational structure of connectionist models is widely believed (see, for example, Anderson 1995, Hinton 1989, Smolensky 1988a) to influence greatly their ability to perform the task for which they designed, whether that task involves learning to categorise stimuli, learning to accurately identify stimuli, or whatever. As Smolensky (1987, cited in Smolensky 1988b) argues:

“a poor representation will often doom the model to failure, and an excessively generous representation may essentially solve the problem in advance” (p. 69).

Hertz, Krogh and Palmer (1991), in a non-psychological context, make essentially this point by demonstrating that the most primitive neural network can learn to solve the most difficult problem, given an appropriate set of pre-processors.

The act, on the part of an experimenter, of abstracting a set of relevant psychological features from a stimulus set constitutes a powerful form of pre-processing. In fact, it is reasonable to suggest that this abstraction makes an inappropriately large contribution to the successful performance of connectionist models employing this type of representation. As Komatsu (1992) argues, the processes of featural selection “bear the bulk of the explanatory burden” (p. 514). Brooks (1991a) identifies the problem more bluntly in asserting:

“this abstraction is the essence of intelligence and the hard part of the problems being solved” (p. 143).

These criticisms appear to be well founded, and certainly cast doubt on the appropriateness of connectionist models of psychological processes built on such psychological featural representational structures. Furthermore, the “generally ... ad hoc” (Smolensky 1988, p. 8) ways in which these features are generated suggests that the entire representational strategy is fraught with

danger. Brooks (1991a) summarises this danger thus:

“it may be the case that our introspective descriptions of our internal representations are completely misleading and quite different from what we really use” (p. 144)

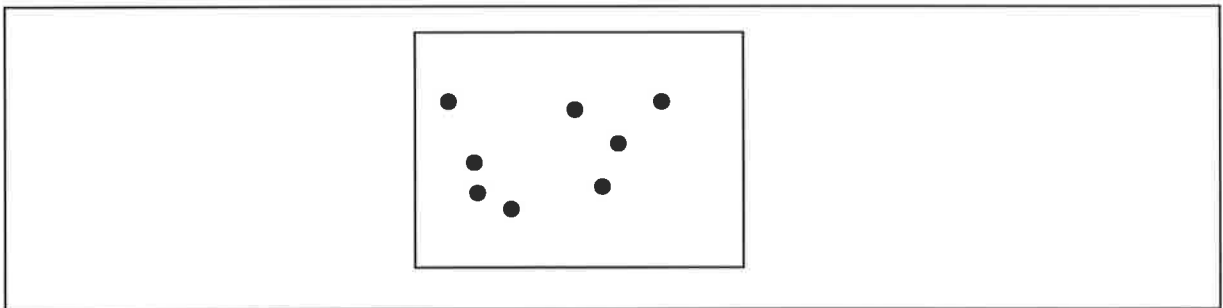
It seems reasonable to require that an approach to accommodating mental structures within connectionist models meet at least three criteria. First, the approach must be explicit - that is, the way in which the representations are derived should be formalised within the model. Secondly, the approach must be objective - that is, the representations that are derived should not depend on who or what implements the model. Thirdly, the approach must be principled - that is, there should be some clear and compelling basis according to which the claim can be made that the model's derived representations are appropriate. The practice of representing stimuli in terms of pre-abstracted psychological features meets none of these requirements.

---

## 2.2. Representation Through Sensory Description

A different approach to realising mental representation in connectionist models of psychological processes is to characterise stimuli in terms of their sensory properties. This approach is particularly amenable to modelling tasks involving simple visual or auditory stimuli - such as line segments or tones - which have been commonly employed in the study of 'low-level' perceptual processes, but which have also been used in the study of 'high-level' cognitive processes.

### 2.2.1. Connectionist Models Employing Sensory Stimulus Representation

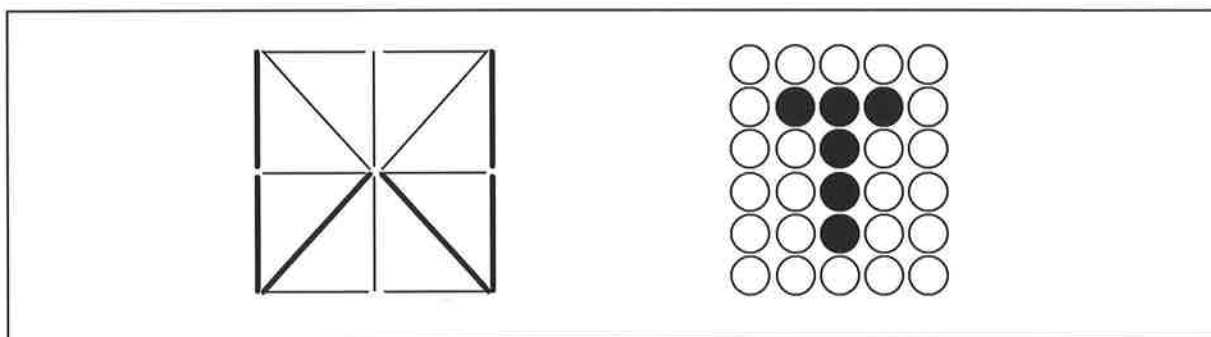


*Figure 2.3. An example of a dot pattern stimulus, of the type employed in Knapp and Anderson (1984). Adapted from Knapp and Anderson (1984), Figure 2.*

A first example of this representational approach is Knapp and Anderson's (1984) connectionist modelling of prototype formation in categorisation. The stimuli encountered by the model take the form of artificial dot patterns, of the type originally described by Posner and Keele (1968, 1970), as depicted in Figure 2.3. These stimuli are used to develop similarity rating and categorisation tasks through which the model is evaluated. Given the highly abstract nature of the stimulus set, the representation of the stimuli is, necessarily, formulated in terms of the physical properties of the stimuli. In particular, the location of all of the dots across the stimulus set is the sole determinant of the model's representational structure.



A second example is found in the widespread connectionist modelling of the psychological process of character recognition. Whilst it is true that some connectionist modelling of character recognition eschews psychological considerations altogether in pursuit of the classificatory accuracy required for practical application (eg. LeCun, Boser, Denker, Henderson, Howard, Hubbard & Jackel 1990), other models either claim or imply some degree of psychological realism in the stimulus representations they employ. For example, the Neocognitron connectionist model (Fukushima 1980, 1988), having been designed as “a network with the same functions and abilities as the brain” (Fukushima 1988, p. 65) could, in its application to character recognition tasks, be considered to belong to this latter class.

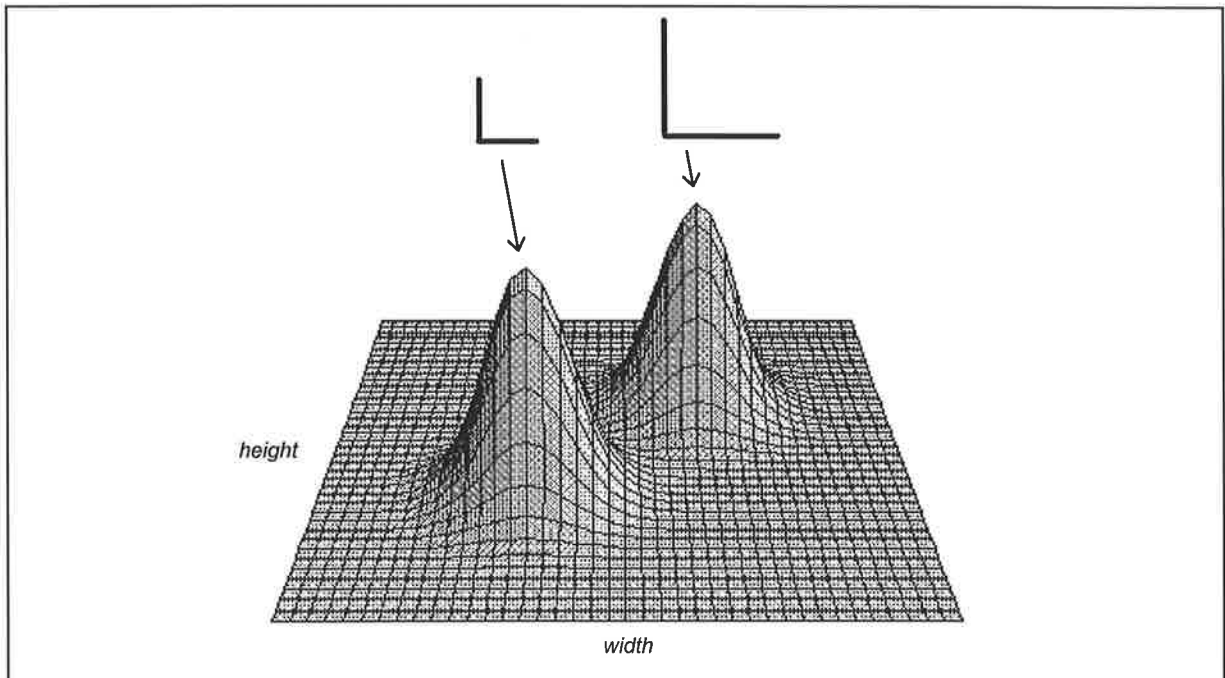


*Figure 2.4. Two commonly employed methods of representing a character stimulus in connectionist models of character recognition.*

As indicated by Figure 2.4, there is some considerable scope with regard to the precise means by which characters can be represented within these models. The letter ‘W’ on the left is constructed from a set of line segments, whilst the letter ‘T’ on the right is composed upon a grid of circular pixels. Typically, each line segment and each pixel would correspond to an input unit of a connectionist model. The important point is that all of the representational structures employed are entirely based upon the physical properties of the character stimuli.

Finally, General Recognition Theory (Ashby & Perrin 1988), although not originally conceived within the connectionist framework, is certainly amenable to connectionist implementation<sup>1</sup>, and provides a clear example of a cognitive process model founded upon a sensory representational structure. In essence, General Recognition Theory represents a stimulus as a probabilistic distribution in a multidimensional ‘perceptual’ space. As a concrete example, consider the simple geometric objects depicted in Figure 2.5, are employed as stimuli in Experiment 1 of Ashby and Perrin (1988). Within General Recognition Theory, such stimuli are represented as a probability distribution in a two dimensional space, where the axes in the space are identified with the length of the line segments.

<sup>1</sup> The feasibility of this reformulation can be traced, for example, by firstly observing the fundamental architectural and procedural similarities between General Recognition Theory and Nosofsky’s (1986) Generalised Context Model (see Ashby & Maddox 1993), and then noting the ease with which a connectionist implementation of the Generalised Context Model has been developed (Nosofsky & Kruschke 1992, p. 213)



*Figure 2.5. The representation of stimuli within General Recognition Theory.*

General Recognition Theory then formalises means by which these stimulus representations may be processed, enabling a model to perform judgments of stimulus similarity, or category membership, or other cognitive tasks. Like Knapp and Anderson's (1984) model, and connectionist models of character recognition, General Recognition Theory is effectively a cognitive process model founded upon stimulus representations which are directly derived from the sensory properties of the stimuli.

### **2.2.2. Evaluating The Sensory Description Representational Approach**

Evaluating the appropriateness of basing representational structure on sensory properties requires a certain degree of subtlety. As a representational approach, it certainly does not suffer from the problem of inappropriate pre-processing inherent in the abstraction of psychological features. Indeed, sensory information is essentially the most primitive and unprocessed information that can be made available to a model from an environment. Neither is the validity of the way in which this information might be described as problematic as it is with regard to psychological features, since well established measurement techniques developed in the physical sciences seem directly applicable. The problem with the sensory representation approach is that, whilst a model must experience its environment to form mental representations of that environment, a convincing argument can be mounted that, in the overwhelming majority of cases, sensory description alone will not serve as an appropriate mental representation of stimuli in that environment.

Rips (1989) provides a partial demonstration of the inadequacies of sensory descriptions as mental representations in arguing that similarity-based cognitive processes operating on sensory representations are incapable of modelling human categorisation. For example, in one experiment, participants are asked whether a circular object, three inches in diameter, is appropriately

categorised as a pizza or a quarter. Not surprisingly, the object tends to be considered to be a pizza despite the fact that its diameter is closer to that of a quarter than that of most, if not all, pizzas. The conclusion Rips (1989) draws from a number of experiments of this type is that similarity-based processes are unable to explain categorisation. This view is one of some considerable contemporary popularity, and is employed as a basis on which to argue for so called 'explanation-based' theories of human conceptual structure (Komatsu 1992, Medin 1989, Smith 1989). An equally valid conclusion to draw from Rips' (1989) experiments, however, may be that it is the sensory representations, rather than the similarity-based categorisation processes, which are inadequate.

In any case, Rips (1989) also argues that sensory descriptions are clearly inappropriate when dealing with 'goal-derived' and 'ad-hoc' categories (Barsalou 1983, 1985) such as 'things to remove from the house if it is on fire'. A similar disassociation between the sensory properties of stimuli and their mental representations has been noted with respect to words (Shepard 1980, Rumelhart & Todd 1993), which have semantic classifications almost entirely unrelated to their perceptual features. Finally, as argued by Medin (1989), the sensory representational approach is clearly completely inadequate if connectionist models are ever to attempt moving beyond representing the external world to accommodate the mental representation of stimuli such as 'ideas', 'emotions', and so on, which do not themselves even have sensory descriptions.

It seems reasonable to suggest that representational structures based on sensory description may be adequate for some modelling, as is evidenced by the notable successes of the three models described in Section 2.2.1. It also seems certain, however, that this sensory representational approach is a limited one, applicable only to a narrow range of stimuli and cognitive processes. As confirmation, consider Knapp and Anderson's (1984) concession that their choice of stimuli is based on the belief that: "Dot patterns seem to be sufficiently unfamiliar and impoverished to be approximated with simple [ie. sensory] representation" (p. 624). Although humans encounter their environment through sensory information, that information must typically undergo a radical upheaval before it becomes a mental representation. As Chalmers, French and Hofstadter (1991) argue:

"Representations are the fruits of perception. In order for raw data to be shaped into a coherent whole, they must go through a process of filtering and organization, yielding a structured representation that can be used by the mind for any number of purposes" (p. 1)

So as not to understate the validity of the three models described earlier, it should be noted that they all incorporate stimulus representations which, although very directly derived from sensory description, are slightly more psychologically sophisticated than the canonical description which might be adopted from the physical sciences. Knapp and Anderson's (1984) model manipulates measures of the physical characteristics of the dot pattern stimuli in order to generate

the continuous analogue of the sum of a set of vector inner products which serves as the actual representational structure on which the model operates. With regard to connectionist character recognition modelling, the interactive activation model developed by McClelland and Rumelhart (1981, see also Rumelhart & McClelland 1982) represents characters not only in terms of sensory line segments, but also accommodates 'top-down' conceptual priming by incorporating word-based representational influences. Finally, Ashby and Perrin (1988) explicitly recognise that the axes of General Recognition Theory's 'perceptual' representation space in the experiment mentioned can be identified with, but will not be identical to, the height and width physical dimensions of the stimulus set. In particular, they expect a monotonic relation between the two sets of dimensions. In this way, the perceptual space representations of General Recognition Theory can accommodate psychological phenomena such as the size-weight illusion.

Nevertheless, it seems reasonable to suggest that the representational structures of all of these models are so tightly linked to sensory description as to be unable to overcome completely the problems raised by Rips (1989) and others (eg. Medin 1989, Smith 1989, see Goldstone 1994, Komatsu 1992 for discussion). It is difficult to imagine exactly how a character recognition model could be extended to be able to categorise words in accordance with their meaning, or how General Recognition Theory could successfully assign a toothbrush and a newspaper to the category 'things to take on vacation' on the basis of any sort of perceptual representation.

Thus, an appropriate conclusion would appear to be that neither of the two most widely adopted representational approaches of connectionist cognitive process models are entirely satisfactory. Therefore, this chapter's survey of previously adopted representational structures in connectionism concludes by examining the approaches of connectionist models which are primarily focused upon the *learning* of mental representations.

---

### 2.3. Connectionist Models Which Learn Internal Representations

Within the connectionist modelling framework, learned internal representations are typically accommodated by the presence of a layer, or layers, of 'hidden' units which indirectly link the input and output units (recall Figure 1.2). Perhaps the simplest example of a connectionist architecture of this type is the 'bottleneck' or 'encoder' architecture (Ackley, Hinton & Sejnowski 1985, Rumelhart & Todd 1993), an archetypal instantiation of which is shown in Figure 2.8. The network is trained, typically using a learning procedure such as backpropagation (Rumelhart, Hinton & Williams 1986), to produce a certain activation pattern across the output layer of units when presented with any one of a set of activation patterns across the input units. By forcing this association to be made through an internal representation layer which has significantly fewer units than both the input and output layers, the network is compelled to form a distributed representation of the input stimuli. If the input and output activation patterns are set to be

identical, the network is referred to as an autoencoder, and effectively realises a principal components analysis of the data contained in the activation patterns (Hertz, Krogh & Palmer 1991). In either case, however, the efficiency of coding demanded of the network's internal representation suggests that such bottlenecks may serve to model some aspect of the representational structure of the domain from which the inputs and outputs are derived.

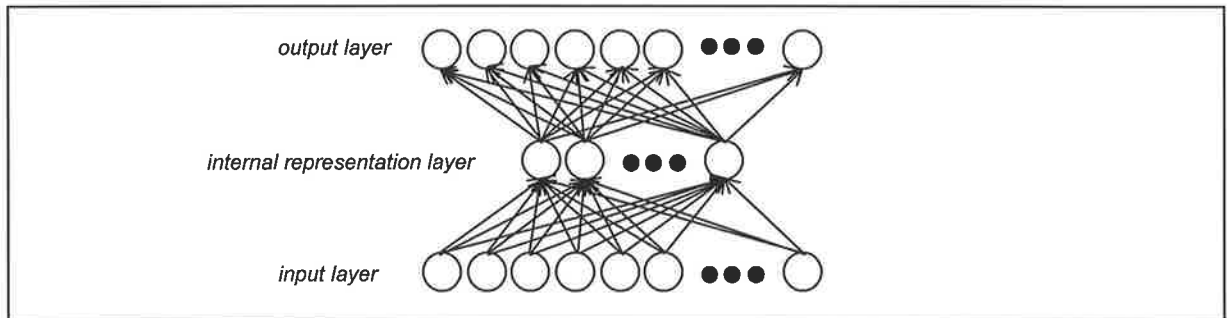


Figure 2.8. The bottleneck or encoder architecture.

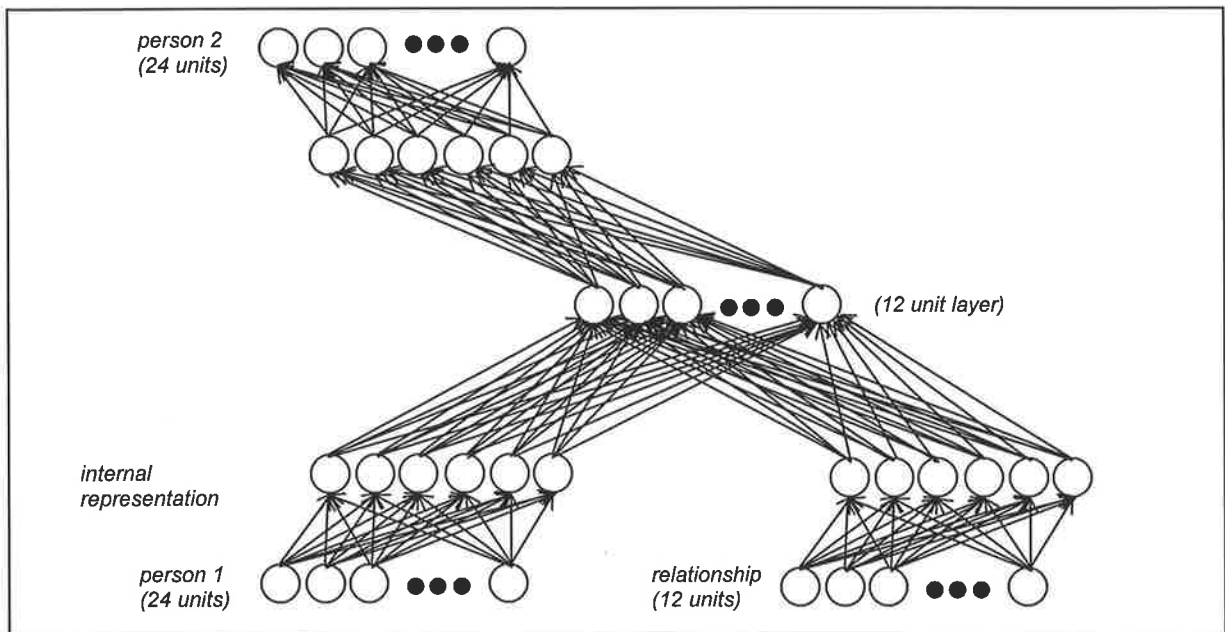
Beyond simple bottlenecks, connectionism affords a considerable degree of modelling flexibility with regard to the exact nature of the internal representations realised by hidden units. The entire network architecture of a model, the learning procedure or procedures employed by a model, and the representations used at the input and output units by a model all influence the internal representations that are learned by that model. Not surprisingly, therefore, previously described connectionist models constitute a plethora of different approaches to learning internal representations. The focus here, however, is on the acquisition of *mental* representation. More surprisingly, perhaps, such a focus significantly limits the models which can be examined. There are few detailed connectionist models which primarily address the way in which mental representations can be learned through the formation of appropriate internal representations.

The two models examined below, the connectionist semantic network and the semantic map, are exceptions in this regard. Both seek the development of internal representations which can, in some way, be justifiably regarded as realising a form of mental representational structure.

### 2.3.1. Connectionist Semantic Networks

Hinton (1989, see also Rumelhart, Hinton & Williams 1986) describes a five layer connectionist model which learns to mentally represent the familial relationships of a group of 24 people from two family trees. The architecture of the network is shown in Figure 2.9. There is a one-to-one correspondence between the 24 people and the 24 units in both the 'person 1' and 'person 2' layers. There is a similar direct relationship between 12 family relationship terms, such as 'has-father', and the 12 units in the relationship layer. On the basis that "the information in a family tree can be expressed as triples of the form (<person 1>, <relationship>, <person 2>)" (Hinton 1989, p. 200), the network is trained, on a set of 100 such triples, by activating the appropriate 'person 1' and 'relationship' units and supplying a teacher value giving the correct

answer at the 'person 2' layer. Any error in the activation levels at the 'person 2' level results in the application of the backpropagation learning procedure<sup>2</sup>, so that, over all 100 triples, the network learns the information contained in the two family trees, and is able to produce the correct 'person 2' answer for four triples not presented during training.



*Figure 2.9. The architecture of Hinton's (1989) model for learning mental representations of family members. Based on Hinton (1989), Figure 5.*

Within the network, the learning of mental representations of the people is accommodated by the bottleneck 6-unit layer which is fully connected to the 'person 1' layer. As was noted above, such bottlenecks force the network to construct a distributed representation of the people, and, to the extent that the individual units correspond to meaningful familial characteristics, this bottleneck layer can be viewed as modelling mental representational structure. The results reported by Hinton (1989) provide some evidence of this occurring. For example, one unit in the bottleneck layer produces activation values consistent with an encoding of the generation of the 'person 1' under consideration.

Rumelhart and Todd (1993, see also McClelland, O'Reilly & McNaughton 1995) extend Hinton's (1989) modelling approach to its natural conclusion in their development of connectionist semantic network models. In essence, these models are connectionist implementations of classic representation systems known as semantic networks (see Collins & Loftus 1975 for an overview). The information captured by semantic nets typically concerns the properties of natural kinds and the inter-relationship between these objects. For example, a semantic network involving plants and animals might be founded on information such as 'an oak is a tree', 'a bird has feathers', 'a fish can swim' and 'a rose is red'. Clearly, the task domains under consideration are broader in scope than the family tree structure used by Hinton (1989). Not surprisingly, therefore, Rumelhart and

<sup>2</sup> In fact, a slightly more complicated learning procedure was employed, in which weights were given a tendency to decay towards zero. This type of approach to optimising the representations learned by the model is considered in Chapter 4.

Todd's (1993) connectionist semantic network, whilst employing essentially the same architectural and learning principles as Hinton's (1989) bottleneck backpropagation network, allows a far more general interpretation of the input and output units in order to accommodate the richness of the traditional task domains of semantic networks.

This re-interpretation is evident in Figure 2.10, which details the generic connectionist semantic network. The input layer which receives information about which stimulus the model is encountering is again connected to a bottleneck layer across which a distributed representation is learned. The structure of this representation is effectively constrained by the relationships of the stimulus to other stimuli and its qualities, capabilities, consequences and other general properties. Specifically, the internal representations are generated by applying backpropagation to modify the connection weights in such a way that the network correctly completes a number of stimulus-relationship input pairs with the appropriate stimulus, property, quality, and so on, as required.

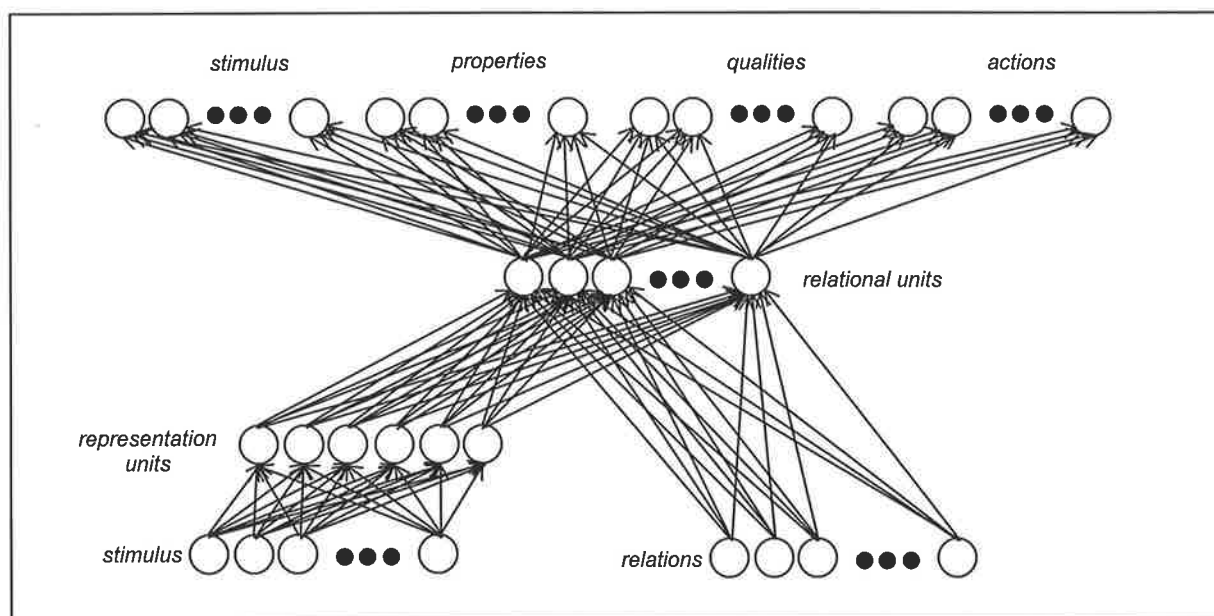
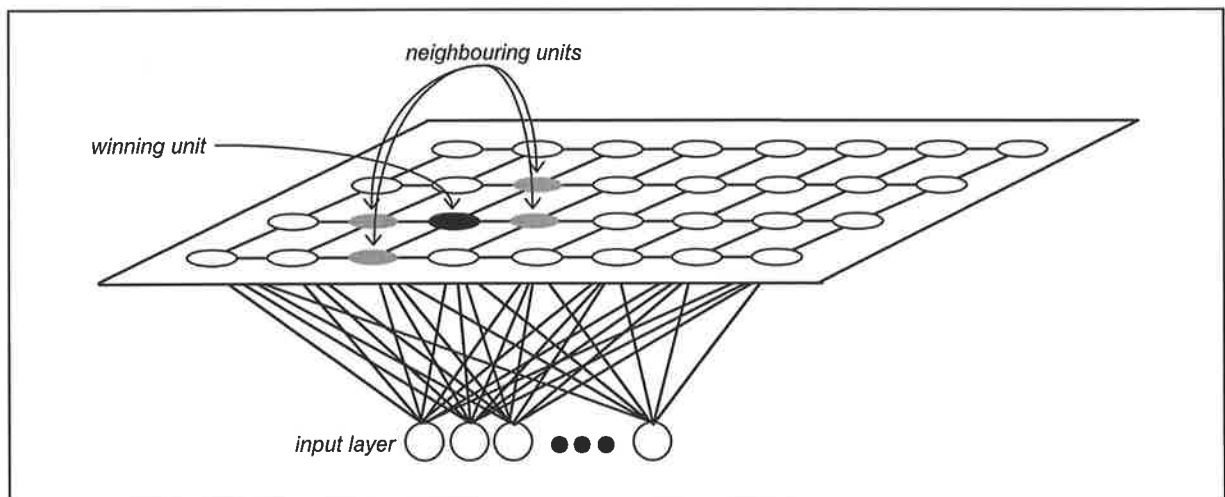


Figure 2.10. The generic connectionist semantic network. Adapted from Rumelhart and Todd (1993), Figure 1.9.

As with Hinton's (1989) model, there is some significant basis on which it could be argued that the internal representations learned by a connectionist semantic network model can be considered to be mental representations. Both Rumelhart and Todd (1993) and McClelland, O'Reilly and McNaughton (1995) present analyses of the representational structure developed by models of this type. Not only are substantive and intuitively plausible psychological interpretations given for many of the units in the representational layer, but the application of hierarchical clustering techniques (see Shepard 1980 for an overview) reveals that the learned representational structure incorporates similarity-based psychological hierarchies. Thus, for example, one unit in the representational layer might be interpreted as indicating the size of a given stimulus, whilst the representational activation patterns for 'oak', 'rose', 'plant' might suggest, upon analysis, that the oak and rose belong to the superordinate concept of plant.

### 2.3.2. The Semantic Map

A different approach to the connectionist modelling of mental representational structure is the semantic map (Kohonen 1990, Ritter 1990). Semantic maps are effectively an application of Kohonen's (1982, 1984, 1990) self-organising map connectionist model to the problem of developing mental representational structures. The basic architecture of a self-organising map is shown in Figure 2.11, and consists of a layer of input units, each of which is connected to every unit in a spatially structured internal representational layer. This spatial structure takes the form of a local, topological ordering whereby a 'neighbourhood' function is defined which establishes, for each unit in the internal representational layer, a set of neighbouring units. Typically, the size of the actual neighbourhood defined by the neighbourhood function decreases during a network's training, with only the topological nature of this neighbourhood remaining constant. Figure 2.11 depicts the commonly employed 'rectangular lattice' neighbourhood function, in which each unit's neighbours are those located spatially one unit away either vertically or horizontally in a two dimensional grid. Other neighbourhood functions, such as linear linkages, or hexagonal lattices, which impart different topologies upon internal representational layers of different dimensionality, have also been employed.



*Figure 2.11. The architecture of a self-organising map.*

The self-organising map's learning procedure involves the adjustment of the weight vectors associated with each internal representational unit, which are of a dimensionality determined by the number of units in the input layer. When information is presented through the setting of activation values at the input layer, a competitive process establishes which internal representation unit's weight vector is, in some sense, the most similar to this input information. The weight vector of this 'winning' unit, as well as the weight vectors of each of the units in its current neighbourhood, are then adjusted by a learning rule so as to match the input more closely. In this way, the self-organising map creates a fundamentally topological structuring of the input information it receives.



The architectures and learning procedures of the self-organising map are entirely replicated in the semantic map. The application of the self-organising map to the problem of developing a mental representational structure of a set of stimuli, therefore, hinges solely on the form of the input information describing these stimuli. In practice, semantic maps are provided with information at the input layer which continually presents stimuli in context, meaning that the mental or semantic similarity which emerges in the internal representation is effectively one which has been equated with contextual similarity.

As a concrete example, consider the semantic map of a set of words described by Kohonen (1990). The input information provided to the model takes the form of coded representations - which are essentially arbitrary and certainly do not specify any form of psychological structure - of every grammatically and semantically sensible three word sentence which can be generated from the set of words. During the learning process, therefore, sentences such as “dog drinks water” and “dog drinks beer” afford the words “water” and “beer” a degree of semantic similarity in accordance with their common neighbouring words. As such, in the stable state of a thoroughly trained network, these semantically similar words will tend to correspond to ‘winning’ units in the internal representational map which are topologically near each other. Results reported by Kohonen (1990) for a set of 30 words using a 10x15 two dimensional semantic map are shown in Figure 2.12. The added partitions indicate the semantic grouping of words in the map in terms of their grammatical classification as nouns, verbs and adverbs. Furthermore, the shaded regions discern additional semantically-based arrangements within the map in which nouns are divided into people, animals, and food and drink classifications. Similar additional arrangements could probably also be suggested in other areas of the map.

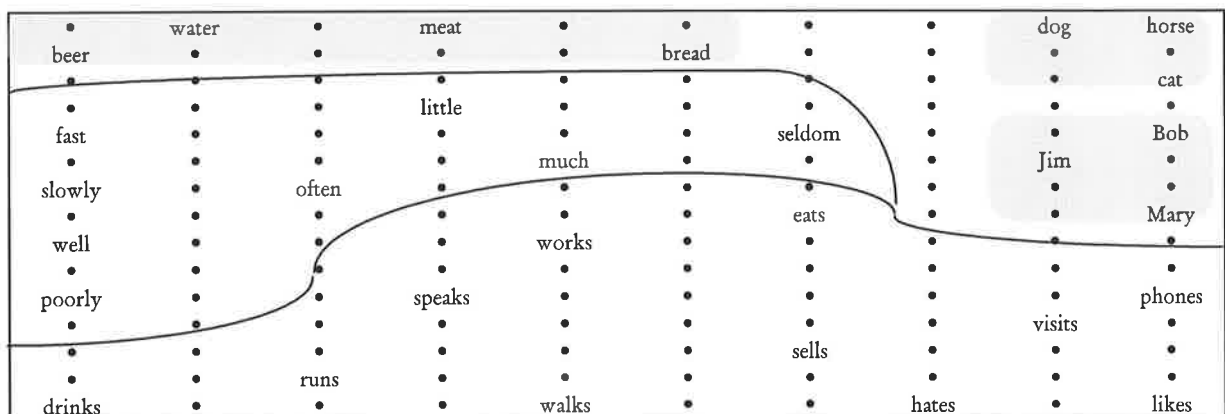


Figure 2.12. The mental representational structure of a 15x10 semantic map. Adapted from Kohonen (1990), Figure 12.

### 2.3.3. Evaluating The Connectionist Models

The featural and sensory description approaches were criticised in Sections 2.1 and 2.2 as, respectively, enacting inappropriate and incomplete theories of mental representation. In contrast, both connectionist semantic networks and semantic maps suffer from the fact that they seem

accepting of virtually any representational structure. Put simply, the connectionist models described above learn internal representations which are insufficiently constrained by psychological theory to be considered as models of human mental representation.

Within connectionist semantic networks, 'mental' representations essentially emerge from the representational constraints imposed by the learning of a set of input/output pairings. It is tempting, therefore, to dismiss this type of modelling on the grounds that any given set of input-output pairings could be accommodated by any number of internal representations. Such a dismissal would be simplistic, given results from mathematical systems theory regarding the existence, amongst all of these possible internal representations, of a unique canonical representation which has some legitimate claim to being *the* appropriate mediating representation of the input/output pairings (see, for example, Casti 1992b). Nevertheless, the internal representations of connectionist semantic networks do not seem to be constructed in a sufficiently principled way. In particular, their network architectures, apart from the inclusion of bottleneck layers, are intuitively appealing, but essentially ad-hoc, constructions designed to accommodate the specific form of the input and output information found in the task domain. The final representational structure developed by a connectionist semantic network is highly dependent upon many features of the network - such as the number of layers, the number of units in each layer, the interconnection of the units, the activation function, and so on - which do not appear to be sufficiently constrained by an analysis of desired representational outcomes. Thus, there may well be network formulations involving bottleneck hidden layers, other than those shown in Figures 2.8, 2.9, and 2.10, which can more readily be regarded as realising mental representations.

Initially, it seems reasonable to suggest that the representational deficiencies of connectionist semantic networks could partially be remedied by the application of established methods from general connectionist modelling which focus on generating internal representations by more direct means. For example, learning processes could be introduced which operate directly on internal representation layers, rather than indirectly constraining these layers by training a network on a set of input/output pairings (see, for example, Lengellé & Denœux 1996). Furthermore, the connectionist semantic network models could be endowed with the ability to modify their network architectures in accordance with the mental representations they accommodate. There is significant non-psychological precedent for this type of approach (eg. Weigend, Rumelhart & Huberman, 1991, see Ash & Cottrell 1995, Haykin 1994, pp. 207-209, Hertz, Krogh & Palmer 1991, p. 158, Reed & Marks, 1995 for overviews). Such prospective solutions, however, effectively amount to more focused re-statements of the fundamental weakness of connectionist semantic network models, namely, that they are not based on an explicit and detailed theory of mental representation. It is precisely such a theory which is required to develop an appropriate learning rule for the internal representation layers, or to describe the mechanisms which alter a model's

network architecture.

A similar situation is evident in relation to the semantic map's property of placing psychologically or semantically similar stimuli near each other in the internal representation layer. Whilst this practice clearly does constitute the enactment of a mental representational principle of sorts, it is not sufficiently developed to generate the type of constraints on the learned representations which would allow the final representational structure to be considered a 'mental' structure. For example, no attempt is made to describe a quantitative relationship between psychological similarity and topological similarity in the internal representational layer, nor have the relative merits of different neighbourhood topologies been articulated. In addition, as noted by Flexer (1996), the quantised nature of the internal representations learned by self-organising and semantic maps limits the resolution with which mental representational structures could be specified.

Furthermore, Bezdek and Pal (1995a) demonstrate that many of the details of the semantic map described by Kohonen (1990) are essentially unimportant in the generation of the final representational structures. This suggests that further analysis or refinement is unlikely to be particularly helpful. It appears that the semantic map's explanatory burden is entirely borne by its similarity-based representational principle. Whilst this principle is not inappropriate, it is neither sufficiently explicit nor sufficiently detailed to be considered a complete theory of mental representation. Consequently the representational structures learned by semantic maps, as with connectionist semantic networks, could not be considered to constitute appropriate models of human mental representation.

#### **2.3.4. Conclusion**

Thus, the conclusion to be drawn from this chapter's survey of previously developed connectionist approaches to modelling mental representations is that neither representations based on pre-abstracted psychological features, nor on sensory descriptions, are appropriate, and that models which learn mental representations tend to do so without recourse to any detailed theory of mental representation. Clearly, therefore, progress in the development of both cognitive process connectionist models, and those models which learn mental representations, requires an appropriate theory of the structure of human mental representations. Chapter 3 proposes for this role the 'psychological space' representational construct.

## Chapter 3: Psychological Space

This chapter presents the ‘psychological space’ approach to the modelling of mental representation developed by Shepard (1957, 1958b, 1987a, 1987b, 1994) which is adopted for the remainder of this thesis. First, the representational concept of psychological space is described and evaluated, and means by which these spaces can be constructed are outlined. Secondly, a number of connectionist models which employ psychological space representations are examined.

---

### 3.1. Foundations Of Psychological Space

The psychological space representational construct is best understood as an attempt to approach the modelling of human cognition in the same way that the physical sciences approach the modelling of natural physical phenomena. Under this view, cognition is conceived of as an array of human mental phenomena which can be explained and predicted through their encoding in an abstract representational space, known as psychological space. Such abstract representations form the basis for many models in the physical sciences. For example, the physical state of an enclosed homogenous gas can be represented by a point in a three-dimensional coordinate space measuring the pressure, volume and temperature of the gas (Casti 1992a, p. 2). In the same way, the psychological space construct attempts to provide a means for the psychological sciences to model human mental representation in an appropriately formulated abstract coordinate space. Moreover, just as the pressure, volume, and temperature of a gas at equilibrium are not independent but, rather, conform to the ideal gas law, the possibility arises of the development of general psychological laws in relation to psychological space.

Typically, abstract physical parameter spaces are structured around quantities such as temperature, mass, and length which closely correspond to sensory descriptions of the type considered in the Section 2.2.1. The previous rejection of these sensory descriptions as mental representations, therefore, amounts to the preclusion of physical parameter spaces as candidate psychological spaces. The basis for this rejection is, in essence, that whilst the representational structure of physical space is appropriate when considering physical phenomena, it is inappropriate with regard to a large number of psychological phenomena. For example, consider a measure of the physical similarity of two objects, generated by a comparison of their representations in physical space. This physical similarity predicts and explains quantitative measures of various physical phenomena with regard to these objects. In fact, this is the basis of the representational structure of physical space: objects are partially represented in terms of, say, their mass to enable the prediction and explanation of their gravitational behaviour, and objects with similar masses are similarly represented and display similar gravitational behaviour. Unfortunately, these physical representations tend not to exhibit the same predictive and explanatory powers in relation to

psychological phenomena. The cognitive ability of humans to remember a list of words denoting objects is largely unrelated to the mass, length, colour, temperature, pressure, and so on, of either these words or the objects themselves. Put simply, physical similarity and psychological similarity imply different abstract representational structures.

The construction of psychological spaces, therefore, is founded upon the notion of representing stimuli to reflect their psychological similarity in relation to psychological phenomena. Shepard (1987b) argues that the most important of all such psychological phenomena is that of generalisation: the cognitive act of behaving (not necessarily overtly) towards one stimulus as if it were another, despite the possible possession of the sensory acuity to discriminate between the two stimuli. This argument is compelling, since all psychological phenomena must be considered in the context of an understanding of their operation under altered conditions, and this understanding must, ultimately, be founded upon an understanding of the phenomena of generalisation.

Moreover, the adoption of generalisation as the most fundamental of cognitive acts renders tenable the construction of psychological space. This construction becomes possible because the abstract notion of psychological similarity which underpins the formation of psychological space is now assumed to dictate, and therefore be implicit in, measures of the cognitive process of generalisation. That is, observed measures of generalisation behaviour, across a stimulus set, are assumed to reflect those patterns of psychological similarity across stimuli which determine psychological space representations, in the same way that physical similarities characterise physical space representations.

### **3.1.1. The Construction Of Psychological Spaces By Multidimensional Scaling**

This approach to the construction of psychological spaces is realised by algorithms which implement the family of statistical techniques known as multidimensional scaling (see, for example, Borg & Lingoes 1987, Coombs 1958, Cox & Cox 1994, Hofmann & Buhmann 1994, Shepard 1980). In relation to a particular set of stimuli, multidimensional scaling operates upon generalisation-based experimental data, such as confusion matrices or similarity ratings, and represents each stimulus as a point in a coordinate space. This space is constructed to be of arbitrary, but minimal, dimensionality, such that the psychological similarity of two stimuli, as measured by the data, is a monotonically decreasing function of the distance between their representative points. In this way, stimuli which are more similar with regards the crucial cognitive process of generalisation are afforded more similar representations.

As an example of the construction of a psychological space by multidimensional scaling, consider a similarity matrix generated by an experimental task requiring the rating of the similarity of each possible pair of a set spectral colours corresponding to the colour labels 'red', 'yellow',

'blue', 'green' and 'violet'. Introspectively, it seems reasonable to expect red and violet colours to be considered more similar than red and green colours, and this intuition is supported by empirical evidence (eg. Ekman 1954, Fillenbaum & Rappenport 1971). Note, however, that the wavelengths which represent the colours red, violet and green in the physical sciences do not predict this pattern of similarity, as the red and violet wavelengths are at opposite ends of the visible spectrum. To accommodate the entire pattern of similarities captured by the confusion matrix, therefore, the application of multidimensional scaling creates a two-dimensional psychological space in which the one-dimensional wavelength spectrum is 'bent' to form a 'horseshoe' or 'colour circle', as depicted for the five colour labels in the left panel of Figure 3.1. This means that the points representing red and violet in psychological space are closer than the points representing red and green. Indeed, in this two-dimensional psychological representation, the distances between all points are monotonically related to the similarities of the colours they represent, and it is these similarities which are presumed to be responsible for the observed confusion behaviour.

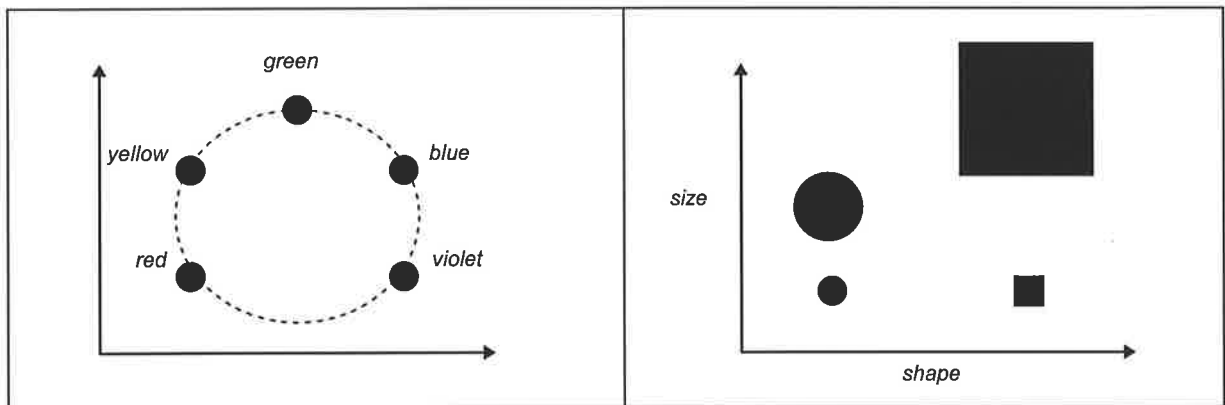


Figure 3.1. Two psychological space representations.

A second example of a psychological space derived through multidimensional scaling is shown in the right panel of Figure 3.1. In this example, the stimulus set consists of four geometric objects which vary with respect to both size and shape. The psychological space representation generated by the application of multidimensional scaling to measures of the similarity of these stimuli is readily interpretable in terms of shape and size dimensions. Note, however, that these dimensions are only implicit in the similarity measures, and are revealed by the multidimensional scaling algorithm seeking a configuration in which inter-stimulus distances monotonically decrease with respect to inter-stimulus similarity. Attainment of such monotonicity is the structural principle which drives multidimensional scaling techniques.

This characterisation of multidimensional scaling, however, would seem to imply that some precise quantitative relationship between psychological similarity and distance in psychological space must be assumed before a quantitatively precise psychological space can be derived. As such, the entire psychological space representational process appears to risk circularity and vacuity. It seems inappropriate to define psychological distances in terms of psychological similarities (implicit

in the generalisation data), construct a representational space on the basis of these distances, and then employ these representations for the prediction of cognitive phenomena which are determined by the patterns of similarities. Lashley (1942, cited in Gregson 1975) articulates this criticism by asserting that explaining generalisation in terms of similarity:

“simply begs the question of the generalizing process, since it assumes that generalization is a function of similarity, whereas similarity is an unexplained result of generalisation” (p. 93)

Multidimensional scaling does appear to do exactly this: psychological similarities are predicted and explained in terms of similar psychological space representations, but these representations are determined through generalisation data which is assumed to be determined by the psychological similarities themselves.

### 3.1.2. The Universal Law Of Generalization

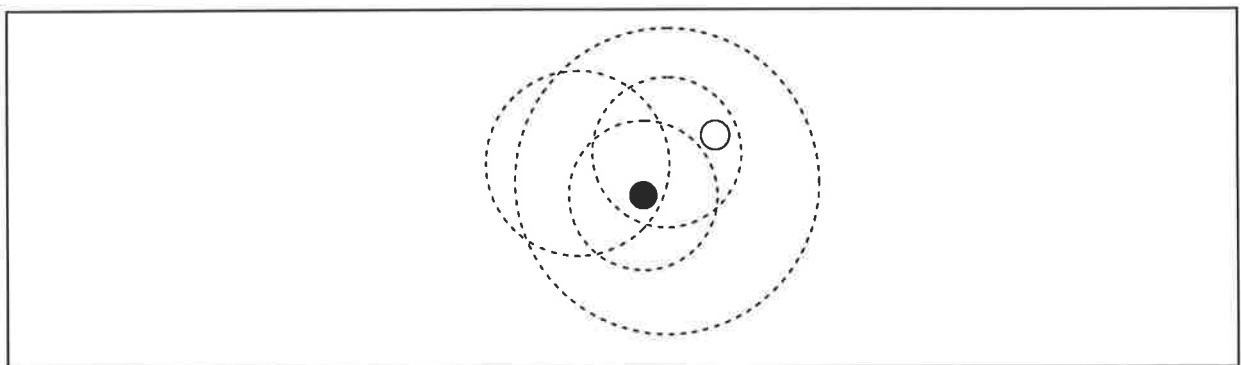
As Shepard (1987a, 1987b, 1988b) argues, however, this difficulty is circumvented by a result referred to as the ‘Universal Law of Generalization’ which specifies an invariant relationship between the psychological similarity of an arbitrary pair of stimuli and the distance between these stimuli in psychological space. Specifically, psychological similarity is given as an exponential decay function of distance in psychological space. This result averts the potential circularity involved in generating psychological spaces through multidimensional scaling because it relieves the need for an ad hoc assumption to be made which affects the derived representational structure. As noted by Shepard (1958b), there are an infinite number of monotonically decreasing functional forms which satisfy the similarity principle of multidimensional scaling. Clearly, each such function will imply a different psychological space representation for any given stimulus set with an associated similarity matrix. The Universal Law of Generalization, however, provides principled grounds for the choice of an exponentially decaying function and, therefore, significantly contributes to the veracity of the psychological space construct as a theory of mental representation. As such, it is important to outline both its empirical and theoretical derivation.

Non-metric (also referred to as interval) multidimensional scaling algorithms (Shepard 1962a, 1962b, Kruskal 1964a, 1964b), in contrast to metric multidimensional scaling algorithms, essentially construct the metric distance relationships which form psychological spaces from ordinal generalisation data which merely provide a ranking of similarities or confusions across a stimulus set. Thus, no requirement beyond monotonicity is made of the function relating distance in psychological space to psychological similarity. Nevertheless, it has been observed across a wide range of empirical generalisation data that the monotonic function derived by the application of non-metric multidimensional scaling is always closely approximated by an exponential decay function (see Shepard 1987a, Figure 1 for a summary). Importantly, Shepard (1987a) reports that it has been demonstrated that this apparent invariance is not a consequence of constraints implicit in

the actual algorithms<sup>3</sup>, but, rather, reflects a property of the cognitive process of generalisation implicit in the data.

The theoretical basis for the Universal Law of Generalization is provided by Shepard (1987a), and hinges on the probabilistic geometry of structures in psychological space termed 'consequential regions'. At the most general level, consequential regions in psychological space correspond to the sets of stimuli in the external world which form natural kinds. More specifically, a consequential region describes the range of different stimuli which share a particular consequence of importance to an individual - consequences such as poisonous, valuable, heavy, and so on - by forming a connected, convex and centrally symmetric region which encompasses all of the points in psychological space representing stimuli with this consequence.

Shepard (1987a) employs consequential regions to derive the Universal Law of Generalization by focusing upon the situation in which a novel stimulus is encountered, and is found to have an important consequence. In this scenario, the lack of information means that the precise extent and position of the consequential region encompassing the point representing the novel stimulus is not known. Rather, the experience of finding a novel stimulus to be consequential is consistent with the existence of a range of potential consequential regions with different degrees of extension. Therefore, the ensuing cognitive process of generalisation involves an evaluation of the probability that other stimuli are encompassed within this range of potential consequential regions. Figure 3.2 depicts this process by showing a black point in psychological space corresponding to the novel stimulus, and a range of surrounding potential consequential regions. The probability that the consequence of this novel stimulus also applies to the stimulus represented by the white point is given by the probability that the white point is encompassed by a consequential region.



*Figure 3.2. Consequential regions in psychological space.*

Clearly, this probability monotonically decreases with respect to distance in psychological space. As the distance between the white point and the black point increases, the probability that a consequential region will encompass both points decreases. The exact quantitative form of the

---

<sup>3</sup> Although there is some evidence (eg. Klock & Buhmann 1997, Goodhill, Simmen & Willshaw 1994) that non-metric multidimensional scaling can impose annular representational configurations, this problem seems limited to binary ordinal data. It is reasonable to assume that psychological space data consists of a larger numbers of ordinal levels.



relationship between the generalisation probability and psychological distance, however, depends upon the assumed prior likelihoods of the possible degrees of extension of the various potential consequential regions. The crux of Shepard's (1987a) derivation is the demonstration that, in fact, this dependence is a very weak one and that, across a number of different distributions, the generalisation probability is closely approximated by an exponential decay function of psychological distance.

Strictly speaking, the method employed for demonstrating this apparent insensitivity relies on the assumption that consequential regions have geometric structures which are both convex and centrally symmetric. Shepard (1987a), however, presents some evidence suggesting the reliance on convexity, at least, may also be weak. In particular, it is shown that the relevant probabilities of encompassment of non-convex, centrally symmetric consequential regions closely approximate those of the convex, centrally symmetric consequential regions employed in the formal derivation.

Interestingly, the argument for an exponential decay relationship between psychological similarity and psychological distance space has also been developed through a number of other, quite different, approaches. Nosofsky (1984) provides a compelling argument for the exponentially decaying functional relationship in seeking to reconcile the psychological space representations constructed by multidimensional scaling with the empirically successful Context Model of categorisation developed by Medin and Schaffer (1978). This argument utilises the fact the exponential decay function is uniquely able to accommodate the Context Model's multiplicative rule for calculating stimulus similarity, given the additive manner in which distances in psychological space are determined. Shepard (1958b, see also Shepard 1990a, Staddon & Reid 1990) derives an exponentially decaying gradient of generalisation on the basis of the 'diffusion' properties of 'stimulus trace' model in which the cognitive effects of the presentation of a stimulus simultaneously spread through psychological space as their strength decays - although this approach has been criticised (Krantz 1967, see also Gregson 1975, p. 103). Finally, support for the Universal Law of Generalization exists in the independent proposal of an exponential decay function to determine stimulus similarity given by Knapp and Anderson (1984), although this evidence could also be criticised, in this case on the grounds that the representational structure employed is not consistent with the psychological space representational construct.

### **3.1.3. Distance Metrics In Psychological Space**

To this point, reference has been made to measures of psychological distance without specifying the metric by which such distances are determined. In fact, the appropriate distance metric depends upon properties of the stimulus set, in a manner which serves to reinforce the veracity of the psychological space representational construct.

Typically, psychological spaces are constructed in coordinate spaces employing one of the

family of Minkowski  $r$ -metrics, which define the distance in an  $N$ -dimensional space between points  $\mathbf{x} = (x_1, \dots, x_N)$  and  $\mathbf{y} = (y_1, \dots, y_N)$  to be:

$$\|\mathbf{x}, \mathbf{y}\|_r = \left( \sum_{i=1}^N |x_i - y_i|^r \right)^{\frac{1}{r}} \quad (3.1)$$

although Shepard (1974) provides an insightful discussion of the possibility of psychological representation in other types of metric or semi-metric spaces, and multidimensional scaling algorithms have been developed (eg. Lindman & Caelli 1978, Cox & Cox 1991) which operate in non-Minkowskian spaces. Nevertheless, the multidimensional scaling techniques by which psychological spaces are commonly constructed usually only accommodate Minkowski metrics (eg. Kruskal 1964a).

The particular Minkowski metrics most often used correspond to the familiar Euclidean ( $r = 2$ ) and City-block ( $r = 1$ ) distance metrics because of their respective association with ‘integral’ and ‘separable’ stimuli (Garner 1974, Nosofsky 1992, Shepard 1991). Integral stimuli are those, such as colours, which are relatively unanalyzable, in the sense that they are not readily perceived in terms of their component dimensions. Separable stimuli, in contrast, are those in which a number of component dimensions can be considered independently, such as the set of geometric stimuli varying in size and shape considered earlier.

Figure 3.3 demonstrates this relationship between distance metrics in psychological space and the nature of the stimuli represented in the space, by indicating the appropriate metric operating within the sample psychological spaces shown in Figure 3.1. The distance between the points representing the red and blue coloured stimuli is Euclidean, whilst the distance between the small circle and large square is measured using the City-block metric.

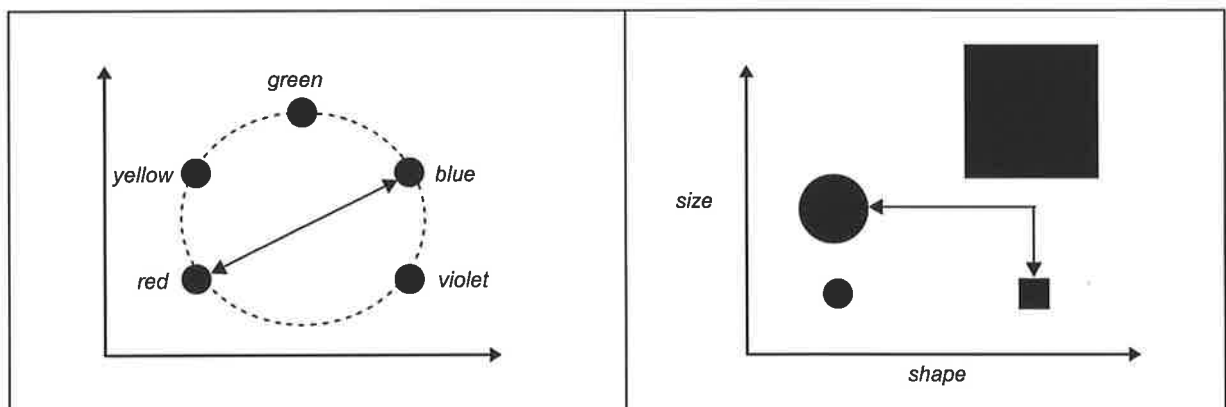


Figure 3.3. Distance metrics in psychological space.

Shepard (1987a) demonstrates that consequential regions, the theoretical tool employed to derive the Universal Law of Generalization, can naturally be extended to account for the operation of different Minkowskian distance metrics in psychological space. Consequential regions with completely uncorrelated degrees of potential extension give rise to City-block metrics, whilst

perfectly correlated ones imply the Euclidean metric. These results are appealing, since the presence or absence of a correlation of component dimensions seems closely related to notions of stimulus separability and integrality. Separable stimuli, for example, are precisely those stimuli with component dimensions which can be considered independently. Between the extremes of perfect correlation and no correlation, where the degrees of extension of consequential regions along orthogonal axes in psychological space are partially correlated, Minkowski  $r$ -metrics with  $r$  values between one and two are realised. It has been suggested (see Shepard 1991, p. 61 for a list of references) that the separable/integral distinction may represent the endpoints of a continuum rather than a dichotomy. If this is the case then some stimulus sets, at least in some circumstances, may be appropriately modelled in a psychological space employing a distance metric with an  $r$  value between one and two. Shepard (1987a, 1991, see also Gati & Tversky 1982) also discusses the possibility of some stimulus sets in which the components of the stimuli 'compete' for attention being modelled by consequential regions with negatively correlated degrees of extension, corresponding to  $r$  values less than one, whilst noting that in this case the distance function is no longer metric because the triangle inequality is violated.

#### **3.1.4. Empirical Evaluations Of Psychological Space Representations**

Empirically, the appropriateness of psychological space representations has been demonstrated by evaluations involving a comparison of the performance of cognitive process models, founded on multidimensionally scaled representations, with data gathered from human performance on equivalent experimental tasks (see, for example, Getty, Swets, Swets & Green 1979, Nosofsky 1984, 1986, 1988a, 1988b). Typically, the tasks involved concern the identification, recognition or categorisation of a set of stimuli. Identification and recognition processes are implemented using the biased similarity choice model (Luce 1963, Shepard 1957), which transforms psychological space representations into identification responses. The modelling of categorisation processes is achieved by employing an extension of the similarity choice model, implicit in the work of Getty et. al. (1979) and explicitly developed in Nosofsky's (1984, 1986) Generalized Context Model, which incorporates the effects of selective attention to the components of separable stimuli.

Inevitably, these empirical evaluations do not constitute conclusive demonstrations that psychological space representations are entirely adequate mental representations. It is, however, fair to conclude that these studies consistently demonstrate an impressive ability of models founded on psychological space representations to model human performance across a range of important psychological phenomena.

More importantly, appending these empirical successes to arguments made earlier for the psychological space representational approach finalises a compelling case: The initial notion that

mental representations are appropriately embedded in an abstract psychological space constructed explicitly for the explanation and prediction of psychological phenomena is an appealing one; the empirical indication of the Universal Law of Generalization, and its subsequent theoretical derivation significantly strengthen this appeal by suggesting that this approach may yield a means of modelling mental representational invariants; the close and natural correspondence between psychological space's distance metrics and the established understanding of the analyzability of component dimensions of different stimuli provides further impetus for adopting psychological space as a theory of mental representation. Empirical demonstrations of the success of cognitive models founded on these representations serve to complete an impressive argument.

### 3.1.5. Criticisms Of The Psychological Space Position

Before adopting the psychological space model of mental representation, however, several criticisms need to be addressed. The majority of these challenges are founded on limitations inherent in the geometric approach adopted by the psychological space theory. More specifically, each of the axioms which define a metric space have been reported to be empirically violated, as have theoretic limitations on the number of possible 'nearest neighbours' in psychological spaces. In addition, the validity of method by which the dimensionality of psychological spaces are usually found has been challenged, casting doubts upon the validity of psychological space representations in general.

If a distance measure  $d$  underpins a metric space, it may be considered to satisfy the following three axioms:

$$\begin{array}{ll}
 d_{ij} > d_{ii} = 0 & \text{positivity} \\
 d_{ij} = d_{ji} & \text{symmetry} \\
 d_{ij} \leq d_{ik} + d_{jk} & \text{triangle inequality}
 \end{array} \tag{3.2}$$

where  $d_{ij}$  is the distance between two distinct points  $i$  and  $j$ . The positivity axiom, which asserts both that the distance between a point and itself is zero, and that the distance to any other point is strictly positive, has been shown (eg. Tversky 1977) to be violated by experimental data gathered in relation to stochastically varying stimuli. The source of these violations would seem to reside in the fact that stimuli which are inherently 'noisy' in this way are not appropriately represented by a single, fixed point in a psychological space. Rather, such stimuli should be represented by a probabilistic distribution of points in a manner similar to that adopted by General Recognition Theory (recall Figure 2.7).

In fact, the implications of noisy stimulus sets for psychological space modelling have been thoroughly considered. Ennis (1988a, 1988b, 1992) extends the Universal Law of Generalization by repeating Shepard's (1987a) derivation under the assumption that stimuli are represented by

Gaussian probability distributions, establishing that, in these circumstances, the invariant generalisation function assumes an inflected Gaussian rather than exponential decay form. This result has been used (see Nosofsky 1988c, 1992, Shepard 1988a, 1988c) to provide a convincing explanation of Nosofsky's (1986) empirical finding that categorisation performance on highly confusable separable stimuli is most accurately modelled by a Gaussian similarity function operating over the Euclidean distance metric.

Extended to stochastically varying, confusable stimulus sets in this way, psychological spaces naturally accommodate experimental data which apparently do not conform to the positivity metric axiom. Allowing momentary fluctuations in the psychological space location of each stimulus, the distance between a particular instance of a stimulus point, and the mean of the distribution of stimulus points from which it is drawn can be non-zero. Similarly, the points representing two distinct stimuli may temporarily coincide.

The symmetry axiom, which asserts that the distance between two points is independent of the direction in which it is measured has also been observed to be violated by experimental data (eg. Gati & Tversky 1982, Tversky 1977, Tversky & Gati 1982). Nosofsky (1992), however, presents a convincing argument that such asymmetries are accommodated by the principled modification of the bias parameters in the choice decision model.

The triangle inequality, which asserts that the distance between two points is less than the sum of the distances from those two points to a distinct third point, is also violated by some similarity data. Although, as noted in Section 3.1.3, this axiom is not necessarily strictly associated with psychological space representations, its empirical violation has been employed (eg. Tversky & Gati 1982) to criticise the psychological space theory of mental representation. Again, however, Nosofsky (1992) argues that the weight of such criticism is substantially alleviated by considering psychological space representations as a part of cognitive process models. In particular, the process of selective attention, formalised in Nosofsky's (1984) Generalized Context Model, which acts to stretch and shrink the component dimensions of an underlying psychological space representation, facilitates the detailed quantitative modelling of the systematic distortions of similarity evident in data which violate the triangle inequality.

Another criticism of psychological space representations relates to the mathematical equivalence or 'quasi-equivalence' of different Minkowski metrics (see Borg & Lingoes 1987, pp. 230-233 for an overview). City-block ( $r = 1$ ) and Ultra-metric ( $r \rightarrow \infty$ ) distance metrics are capable of achieving different, but equally error free, psychological space representations. Similarly, for Minkowski metrics with  $r < 2$ , there is a lawfully related  $r$  value greater than 2 which can accommodate a different representational structure with essentially the same minimum error. These results could be taken to suggest that psychological space representations may be somewhat arbitrary, in the sense that the problem of finding an appropriate distance metric and

representational configuration is underdetermined. Such a criticism, however, does not place a psychological interpretation upon the operation of a distance metric with  $r > 2$ , and it is difficult to conceive of such an interpretation. Pure integrality at  $r = 2$  would seem to constitute a psychological limit upon the degree to which underlying stimulus dimensions can be combined. In contrast, distance metrics corresponding to all  $r$  values  $0 < r \leq 2$  can be given a substantive psychological meaning, as outlined in Section 3.1.3. Thus, it seems reasonable to suggest that the distance metric underpinning a psychological space be restricted to the range  $0 < r \leq 2$ , in which case the problem of metric equivalencies does not arise. Such a restriction would have the advantage of being consistent with Shepard's (1987a) demonstration of the relationship between the correlation of dimensions of consequential regions and the distance metric, since the correlational limits of  $\pm 1$  correspond to the metric parameters  $0 < r \leq 2$ . In fact, the metric equivalency results could perhaps be seen as further impetus for restricting  $0 < r \leq 2$ , since they demonstrate that metrics with  $r > 2$  do not afford further representational possibilities.

Another critique of psychological space representations is provided by Tversky and Hutchinson (1986), who argue that limitations inherent in the geometric representations prevent the appropriate modelling of some conceptual structures. In particular, it is shown that a geometrically imposed upper limit for the number of nearest neighbours of any given representational point must be exceeded to accommodate some similarity data. This difficulty seems most prevalent in the case of conceptual stimulus domains which are amenable to a hierarchical structuring. For example, Tversky and Hutchinson (1986) analyse a stimulus domain consisting of the word 'fruit' and a set of 20 individual fruits words, such as 'lemon', 'orange', and 'banana'. The empirical similarity ratings require that 'fruit' be the geometric nearest neighbour of 18 of the individual fruit words, yet an otherwise appropriate multidimensionally scaled representation locates 'fruit' closest to only two other stimulus points. Even more fundamentally, Tversky and Hutchinson (1986) note that it is, in principle, impossible for a point to be the nearest neighbour to more than 5 points in a two dimensional space, or 11 points in a three dimensional space. Thus, to accommodate the nearest neighbour relationships evident in the similarity ratings, multidimensional scaling solutions would need to be derived in spaces of inappropriately high dimensionality.

The insight provided by the nearest neighbour measure developed by Tversky and Hutchinson (1986) is readily appreciated by considering the judgments involved in the collection of similarity rating across hierarchical stimulus domains of this type. The decision processes involved in judging the similarity of 'orange' and 'lemon', or 'orange' and 'date', would seem to differ fundamentally from those involved in judging the similarity between 'fruit' and 'orange'. The first type of judgment involves only a judgment of similarity, whilst the second judgment essentially requires a determination of the typicality of 'orange' as a member of the superordinate concept

‘fruit’, perhaps through the judgment of similarity to a prototype. Employing the terminology of Pipkin (1982), the first comparison is a point-point comparison, whilst the second is a point-set comparison. In effect, therefore, the analysis of Tversky and Hutchinson (1986) highlights the inability of multidimensional scaling to operate on point-set similarity data. Thus, nearest neighbour measures reveal limitations in the ability of psychological space representations to represent stimuli which exist at different levels of a conceptual hierarchy.

As Shepard (1994) notes, these limitations may be derived from the assumption that consequential regions are connected in the sense that:

“between the points corresponding to any two objects of [a] kind ... there is always a continuous path in the representational space which falls entirely within the consequential region for that kind” (p. 23).

Such a prescription relates to the notion of the ‘basic’ level in conceptual hierarchies (see Mervis & Rosch 1981, Rosch 1978 for overviews), which may be considered as the most general level at which stimuli belonging to a given concept can be perceptually continuously deformed into one another. Consequential regions, therefore, can only model concepts up to the generality represented at the basic level. Concepts which exist at a superordinate level of a conceptual hierarchy, such as those explored by Tversky and Hutchinson (1986), cannot be accommodated by consequential regions, and hence are not amenable to psychological space representation. Whilst the extension of the psychological space theory to accommodate such conceptual structures remains a priority for future research, it should be noted that the critique of Tversky and Hutchinson (1986) does not discredit psychological space representations involving non-hierarchical stimulus domains.

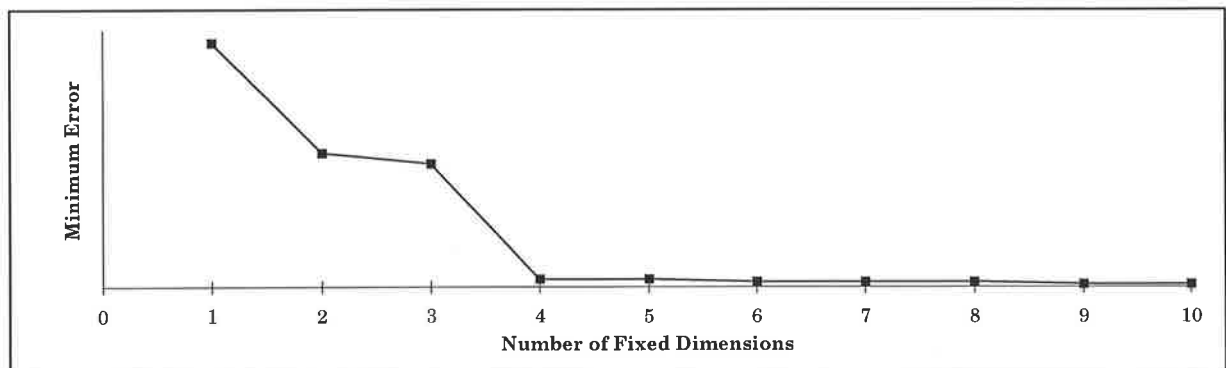


Figure 3.4. An error elbow.

Finally, a sustained criticism of the psychological space theory relates to difficulties in determining the appropriate dimensionality of the derived representational structure. Early research (eg. Kruskal 1964a, Shepard 1962b) suggested that an examination of the minimum error (or ‘stress’) values across representational spaces of increasing dimensionality reveals an error ‘elbow’ of the type illustrated in Figure 3.4. Clearly, the (fictitious) data which are being multidimensionally scaled in Figure 3.4 are best represented in a four dimensional space, since the

addition of the fourth dimension accommodates a significant reduction in error, but the subsequent addition of dimensions does not achieve any significant further reduction in the minimum attainable error. Graphically, this state of affairs is apparent in the sharp reduction of the gradient at the point corresponding to the four dimensional space, coupled with the relatively low absolute value of the error at this point.

Unfortunately, it has been widely noted (eg. Borg & Lingoes, p. 68, Grau & Nelson 1988) that the application of multidimensional scaling to many data sets reveals a gradually decreasing error minimum as dimensionality increases, thus making a decision regarding the 'true' dimensionality of the space subjective and problematic. It is, however, worth noting that the majority of these data sets either consist of psychologically integral stimuli, or are implicitly treated as such through the employment of the Euclidean distance metric. Given that psychological spaces which assume this metric are of intrinsically arbitrary rotation, it seems unlikely that attempts to provide an objective specification of the dimensionality of the space are warranted. Since psychological spaces representing integral stimuli have no substantive dimensional structure, in the sense that component dimensions cannot be given psychological interpretation, the exact number of dimensions which are ultimately included is of little import. The primary function of the derived psychological space is to accommodate the pattern of inter-stimulus similarities given by the data. Thus, psychological spaces employing the Euclidean metric should simply employ the minimum number of dimensions required to achieve a sufficiently low error value, and need not be concerned with the presence or absence of an error elbow.

Psychological spaces which operate under non-Euclidean distance metrics, in contrast, do contain a preferred dimensional structure which must be evident in any derived representation. For example, the application of the process of selective attention to the component dimensions of separable stimuli, as described earlier, relies upon the existence of the appropriate coordinate axes in the representational space. As such, non-Euclidean psychological spaces would seem to require an error elbow for their principled derivation. Evidence that real data does not reveal such an elbow, however, is less than conclusive due to well documented (eg. Arabie 1991, Borg & Lingoes 1987, Hubert, Arabie & Hesson-Mcinnis 1992, Shepard 1974) doubts concerning the ability of many multidimensional scaling algorithms to derive valid psychological spaces under these conditions. Put simply, there are compelling grounds on which to doubt the asserted minimality of error values for non-Euclidean spaces.

Given these doubts, derived error values afford no insight with regard to the presence or absence of error elbows. It remains possible that error elbows are a naturally emergent property of the interplay of the representational principle which underpins multidimensional scaling and the structure of natural kinds in the world from which the stimuli are drawn. This issue remains an open research question, which is further discussed in Section 5.2.1. For the moment, it should be



concluded that purported demonstrations of the absence of error elbows when real data is multidimensionally scaled do not discredit the psychological space model of human mental representation.

Having addressed these criticisms, the remainder of this thesis adopts the position that human mental representation can be appropriately modelled in psychological spaces, and assumes the Universal Law of Generalization and its associated distance metric results. The aim of developing a connectionist model which learns mental internal representations thus becomes the aim of developing a connectionist model which learns psychological space internal representations. An appropriate starting point, therefore, is to examine previously suggested connectionist models which have employed psychological space representations.

### 3.2. Connectionist Models Using Psychological Space Representations

The first two connectionist models employing psychological space representations to be considered are directly interpretable in terms of Shepard's (1987a) description of the psychological space construct. Rather than assuming the Universal Law of Generalization, these models faithfully implement the consequential regions employed in Shepard's (1987a) derivation, allowing the exponential decay function relating psychological similarity to distance in psychological space to emerge epiphenomenally. In this sense, the way in which both models realise the Universal Law of Generalization, described here as the 'consequential region' approach, could also appropriately be characterised as being a 'first principles' approach.

#### 3.2.1. The Consequential Region Model

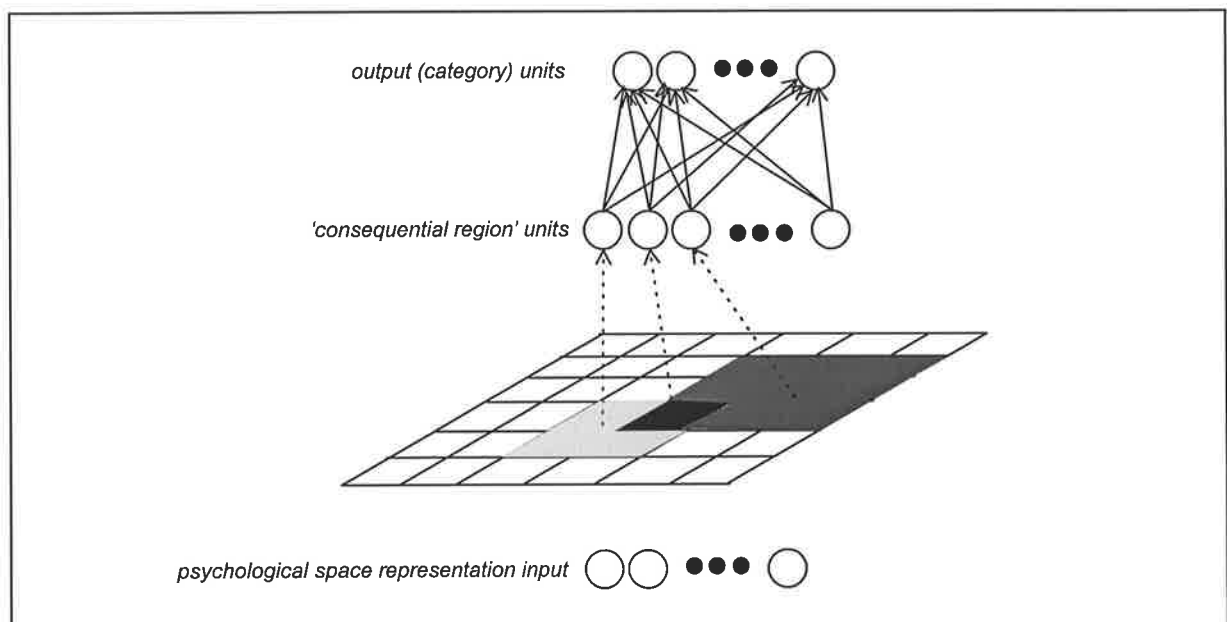


Figure 3.4. The Consequential Region model. Adapted from Shanks and Gluck (1994), Figure 2.

The architecture of the first of these models, the Consequential Region Model (Shanks &

Gluck 1994) is shown in Figure 3.4. The dimensionality of psychological space is pre-determined, and stimuli are presented to the model in pre-determined psychological space coordinates. The model's second layer consists of a set of 'consequential region' units which are constructed in one-to-one correspondence with the potential consequential regions in psychological space, which is quantised through the imposition of a square grid. The compact, convex and centrally symmetric consequential regions thus consist of all possible rectangles in the case of separable stimuli, but are limited to squares when integral stimuli are being presented.

In either case, the presentation of a stimulus activates only those consequential region units which correspond to regions containing the stimulus' representing point in psychological space. The activation of the consequential region units, in turn, activates a layer of output units, generally associated with categorical responses, as mediated by a set of connection weights.

Learning in the Consequential Region Model is restricted to the adjustment of these connection weights by a learning rule, generated by gradient descent on an error derived from appropriate categorical feedback. This learning rule differs from the standard gradient descent approach - also known variously as the 'least mean squares', 'delta', or 'Widrow-Hoff' learning rule (see Anderson 1995) - to the extent that separate learning rate parameters are maintained for the increment and decrement of the connection weights. Whilst Shanks and Gluck (1994) detail the Consequential Region Model's ability to accurately emulate human categorisation performance on a variety of tasks, these demonstrations are limited in scope by the model's inability to incorporate the effects of selective attention, although Shanks and Gluck (1994, pp. 84-85) do suggest briefly how this shortcoming might be remedied.

### **3.2.2. Shepard And Kannappan's Model**

A more sophisticated first principles approach to the connectionist representation of psychological space mental representation is described by Shepard and Kannappan (1991). The architecture of this model is shown in Figure 3.5. The model deals only with sets of uni-dimensional stimuli; that is, with stimuli which can be appropriately represented in a one-dimensional psychological space. The input layer of the model contains units which are in one-to-one correspondence with the elements of the stimulus set. The output of the model is determined by the pattern of activation across a layer of response units. Mediating the flow of information from the stimulus input layer to the response output layer are a pre-determined number of layers containing units which correspond to potential consequential regions within the psychological space. These layers are arranged according to the size of the potential consequential region they represent, with smaller regions being positioned in layers closer to the stimulus input layer.

A stimulus is presented to the model by activating the unit corresponding to that stimulus, and leaving all other units in the input layer inactive. The activation of an input unit is propagated

upwards through the network, serving to activate those consequential region units which lie in the input unit's established 'cone of activation', as shown in Figure 3.5. The magnitude of this activation exponentially decays with respect to the consequential region unit's distance, in terms of layers, from the stimulus input layer. Finally, the pattern of activation across the response units is determined by the activation of both the stimulus input and consequential region units, as mediated by a set of connection weights.

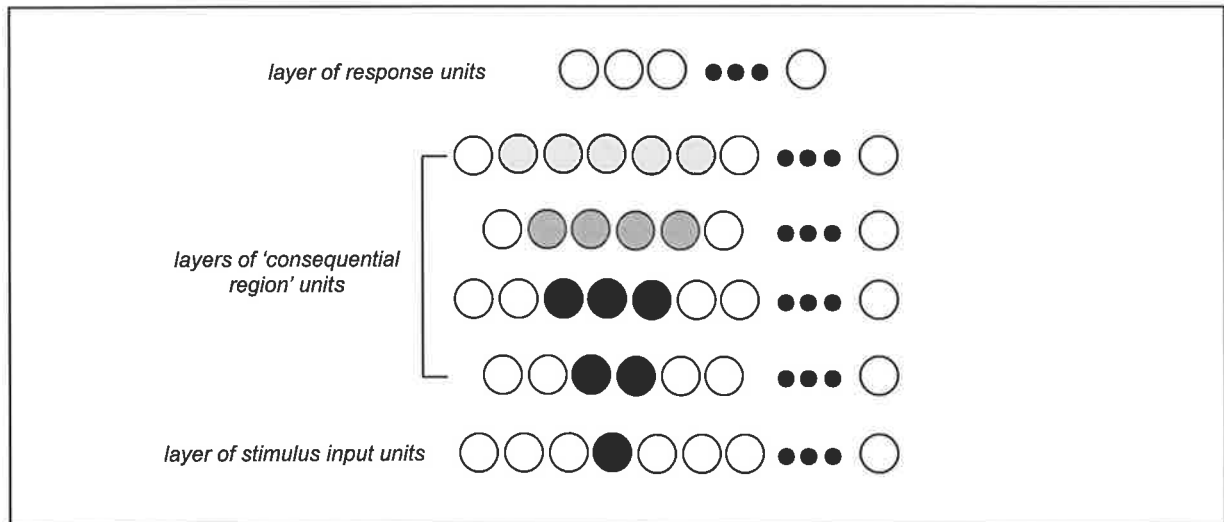


Figure 3.5. Shepard and Kannappan's (1991) generalisation model. Active units lie within the 'cone of activation'.

Learning within this model primarily involves the adjustment of these connection weights. Once again, the learning rule, in effect, employs gradient descent on an error function derived from the reinforcement or non-reinforcement of the response produced to the presented stimulus. Shepard and Kannappan (1991) also describe an extension of the basic model in which connection weights are added between the input and consequential region units, and are adjusted using backpropagation.

Whereas Shanks and Gluck's (1994) Consequential Region Model is primarily applied to category learning tasks, the evaluative focus of Shepard and Kannappan's (1991) model centres on the learned adjustment of gradients of response generalisation for uni-dimensional stimulus sets. The results reported by Shepard and Kannappan (1991) demonstrate that the model captures a range of established properties of generalisation gradients, although they appropriately conclude that "this is just the beginning of the connectionist exploration of the implications of the generalisation theory in more complex cases" (p. 670). In particular, the model's restriction to uni-dimensional stimuli is a severe one, evidently addressed by Shepard and Tenenbaum (1991), although their multidimensional extension of the model remains unpublished (Tenenbaum, personal communication, October 1995).

Nevertheless, it could be argued that Shepard and Kannappan's (1991) interpretation of the consequential region approach has a fundamental advantage over that of Shanks and Gluck's (1994) Consequential Region Model in that the layering of the consequential region units captures what

are presumably important topological properties of psychological space. Effectively, the Consequential Region Model collapses all of the consequential region units which are layered according to size in Shepard and Kannappan's (1991) model into a single layer and, in so doing, discards this potentially useful information. Intuitively, it seems likely that this information could contribute to a model's realisation of psychological space, although the exact nature of this contribution remains to be demonstrated.

### 3.2.3. The Radial Basis Function Approach

A different approach to the connectionist realisation of psychological space is achieved if the Universal Law of Generalization and its associated distance metric results are assumed, and are not required to be continually re-derived by the model. Indeed, if these assumptions are made, a connectionist network with a particular architecture, known as a 'radial basis function' architecture (Lowe 1995, Moody & Darken 1989, Poggio & Girosi 1990), is naturally able to implement psychological spaces.

As mentioned in Chapter 1, connectionist models can accommodate the fundamentally geometric representational structure required by the notion of psychological space. The representation of any point in a coordinate space of any dimensionality can be achieved through the activation of a layer of units in a connectionist network, where there is one unit for each dimension in the space, and the activation value of each unit corresponds to the coordinate value, on that dimension, of the point being represented. This is precisely the way in which a radial basis function network represents input information.

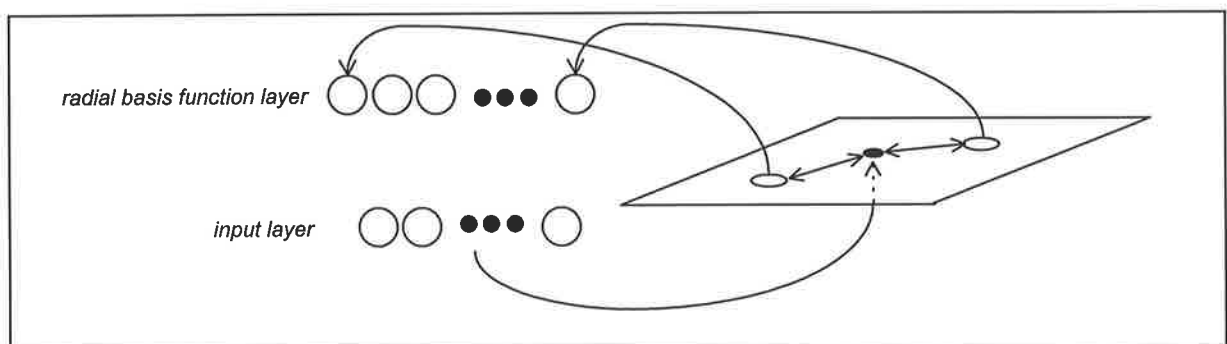


Figure 3.6. Part of the radial basis function architecture.

Furthermore, radial basis function networks assume that each unit in the subsequent layer is positioned in the same coordinate space as the point corresponding to the input information. Once such an input is presented, the activation value of each unit in this subsequent 'radial basis function' layer is determined by measuring the distance between the input point and the unit's location, and then evaluating a fixed basis function, using this distance measure as the independent variable. Figure 3.6 depicts this process by detailing the correspondence between the standard layered interpretation of the network, and its geometrical re-interpretation.

Therefore, within a radial basis function architecture, by adopting a one-to-one correspondence between elements of a stimulus set and units in the radial basis function layer, placing the units in their psychological space positions as determined by multidimensional scaling, ascribing an appropriate Minkowskian distance metric, and specifying the basis function to be one which exponentially decays as distance increases, a connectionist network realises a complete model of the psychological space representational construct.

### 3.2.4. The ALEX Model

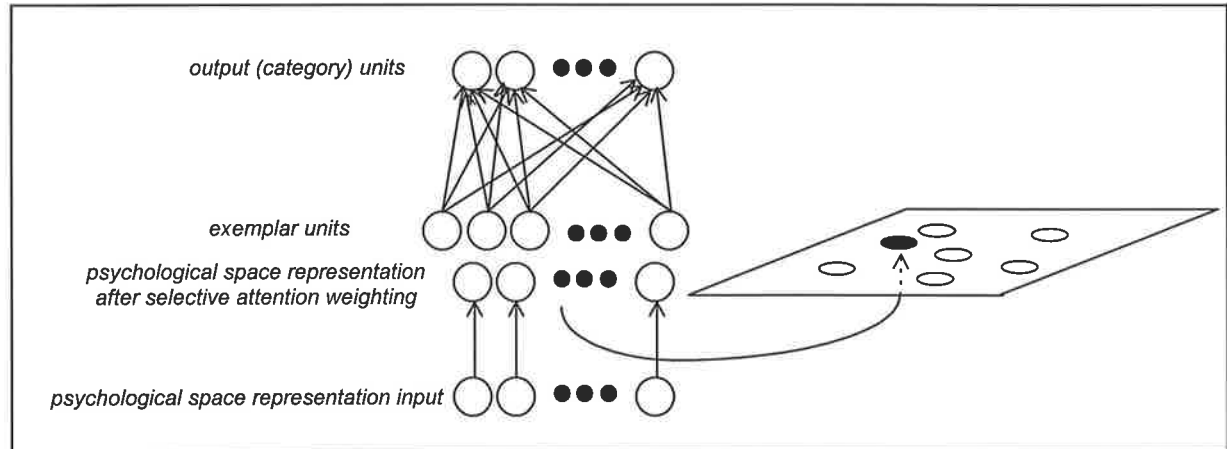


Figure 3.7. The ALEX model.

This approach is the one employed by the ALEX model (Kruschke 1990, 1992, Nosofsky & Kruschke 1992), which is, in essence, a connectionist implementation and extension of Nosofsky's (1984, 1986) Generalized Context Model. The architecture of the ALEX model is shown in Figure 3.7. An element of the stimulus set is presented to the model by setting the activation values of the input layer to the previously derived psychological space coordinates of that stimulus. These coordinates are then transformed by non-negative weightings which model the process of selective attention, effectively scaling the axes of the psychological space. This transformed stimulus representation then activates 'exemplar' units in the radial basis function layer, each of which corresponds to a stimulus, using the computational approach described in Section 3.2.3'. Once activated, the exemplar units generate, through a set of connection weights, a pattern of activation at a layer of output units. Since the ALEX model has been almost exclusively applied to categorisation tasks, the output units are usually identified with categorical associations. In particular, Kruschke (1990, 1992) routinely uses the activations in the output layer to derive response probabilities for each of the various categories to which the stimulus may belong.

Learning in the ALEX model involves both the adjustment of the connection weights linking exemplar units to the category output units, and the adjustment of the selective attention weightings. Both of these adjustments are accomplished by learning rules which perform gradient

descent on the sum squared error between the values of the output units, and a set of ‘teacher’ units which specify the correct categorical association of the presented stimulus.

Empirical evaluations of the ALEX model demonstrate an impressive ability to emulate human performance across a wide range of categorisation related tasks (see Kruschke 1990, 1992, 1993a, 1993b, Nosofsky & Kruschke 1992, Nosofsky, Kruschke & McKinley 1992). These include the seminal attentional category learning task examined by Shepard, Hovland and Jenkins (1961), filtration and condensation tasks (Garner 1974), tasks involving the potential for catastrophic forgetting (Ratcliff 1990), tasks requiring the exhibition of three-stage learning (Rumelhart & McClelland 1986), tasks involving the differentiation between linear and non-linear category boundaries (Medin & Schwanenflugel 1981), and tasks requiring sensitivity to correlated stimulus dimensions (Medin, Altom, Edelson & Freko 1982). The model is also able to successfully account for human performance in a limited subset (Lewandowsky 1995) of those categorisation tasks which require the utilisation of base-rate information (Medin & Edelson 1988). Finally, Choi, McDaniel and Busemeyer (1993) demonstrate that the principled initialisation of the ALEX model’s connection weights results in the impressive emulation of differences in human performance on tasks involving the learning of conjunctive, disjunctive, and a range of other categories formulated through the logical operations of a predicate calculus (Bourne 1974).

### 3.2.5. The ALCOVE Model

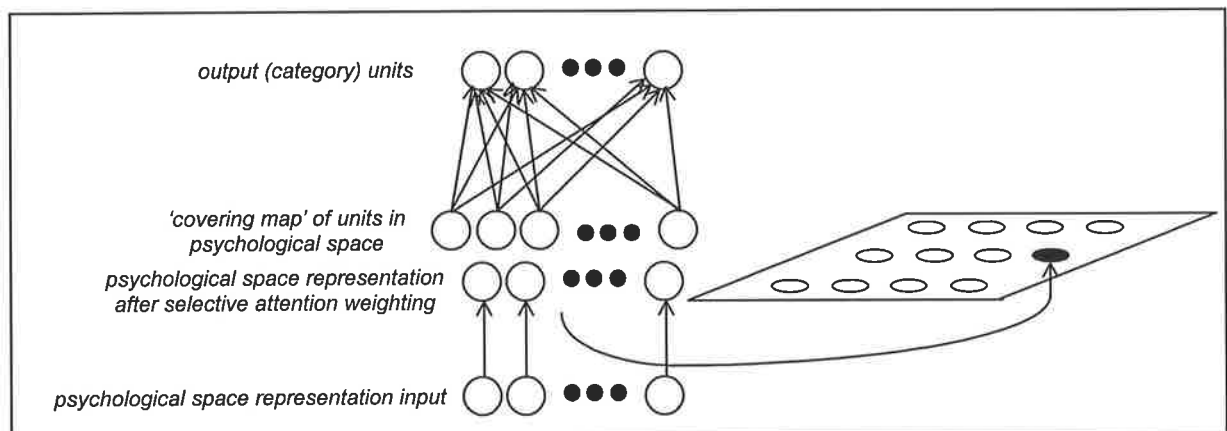


Figure 3.8. The ALCOVE model.

Kruschke (1992), however, observes that the ALEX model’s pre-positioning of exemplar units in psychological space is inappropriate to the extent that “the model cannot assume knowledge of the exemplars [stimuli] before it has been exposed to them” (p. 39). A variant of the ALEX model, known as the ALCOVE model, addresses this weakness by disassociating units in the radial basis function layer from specific stimuli, instead placing a ‘covering map’ of units across psychological space. This covering map is constructed by deciding upon the appropriate

<sup>4</sup> In fact, the applications of ALEX described in Kruschke (1990) employs a Gaussian, rather than exponential, basis function. The results reported in Kruschke (1990) are, however, subsequently replicated (eg. Kruschke 1992) using the exponential decay function.

dimensionality of psychological space on the basis of the multidimensional scaling of the stimulus set, partitioning the space into equal (hyper)cubes in accordance with a density parameter, and then randomly placing a unit within each of these partitions.

In all other architectural and procedural respects, the ALCOVE and ALEX models are identical, as can be seen in the depiction of the ALCOVE model in Figure 3.8. Evidently (see Kruschke 1993a, p. 63, footnote 1), the categorisation performance of the ALCOVE model closely parallels that of the ALEX model, although the ALCOVE model's performance is detailed only by Kruschke (1990).

Again, however, Kruschke (1990, pp. 132-133) observes limitations in the representational structure which realises psychological space. In the case of the ALCOVE model, the exponential increase in the number of covering map units required as the dimensionality of psychological space increases would seem to require potentially unrealistic architectural resources of the network. In addition, it should be noted that the pre-determination of the dimensionality of psychological space is inappropriate.

As a suggested remedy to the requirement of excessive numbers of covering map units, Kruschke (1990) provides some discussion of another form of learning which might be implemented within the ALCOVE model, involving the adjustment of the location of the covering map units in psychological space, to "distribute them over only those regions of the ... [psychological] space in which stimuli occur" (p. 132). These discussions note that those learning techniques identified with self-organising map networks (Kohonen 1982, 1988a, 1990), as previously applied within radial basis function architectures (eg. Moody & Darken 1989), may be more appropriate than those based on gradient descent on error. Finally, again attempting to limit the required number of covering map units, Kruschke (1992) raises the possibility of a mechanism which creates or recruits units in psychological space in accordance with the location of stimuli, and is mediated by some form of novelty parameter. In reported applications of both the ALEX and ALCOVE models, however, the exemplar and covering map units, respectively, are permanently fixed at pre-determined psychological space locations<sup>5</sup>.

### **3.2.6. Comparing The Consequential Region And Radial Basis Function Approaches**

In one sense, it is difficult to compare the consequential region and radial basis function approaches because they differ more in the level at which they attempt to model psychological spaces than in the features of these spaces which they actually model. In this light, the assertion of Shanks and Gluck (1994) that the Consequential Region Model should be viewed as an alternative to, rather than a competitor of, the ALEX and ALCOVE models appears well founded. Essentially, the consequential region approach provides a first principles alternative to the radial

basis function approach for the implementation of the psychological space representational construct.

There are, however, at least three reasons for preferring the radial basis function approach, despite the consequential region approach's close adoption of the theoretical machinery employed by Shepard (1987a). The 'first principles' approach is advantageous only to the extent that it provides the modelling flexibility to capture important features of psychological spaces which are not sufficiently well approximated by the Universal Law of Generalization and its associated results. Whilst Shepard and Kannappan (1991) suggest that the modelling of chronometric cognitive tasks might be accomplished using the consequential region approach and could not be accommodated by the radial basis function approach, it seems reasonable to assert that, at present, such abilities remain to be conclusively demonstrated. On this basis, the parsimonious realisation of the psychological space representational construct provided by the radial basis function approach is to be preferred.

Secondly, the representational resources required by the consequential regions approach, as measured by the number of units in the consequential region layer(s), tends to be prohibitively large. For example, in modelling a 15x21 psychological space grid, the Consequential Region Model employs 33,264 consequential region units. Whilst Shanks and Gluck (1994) do explore the possibility of not establishing a consequential region unit for every potential consequential region, which clearly reduces the architectural and processing demands on the model, it seems likely that this economising would concurrently serve to degrade the emergent generalisation gradient's approximation to an exponential decay function. In any case, the number of units required in the consequential region layer increases rapidly as the number of quantised cells in psychological space is increased. This situation arises when either the bounds on psychological space are extended, or the resolution within the space is improved. Much the same type of criticism could be directed at Shepard and Kannappan's (1991) model, particularly with regard its multidimensional extension.

Furthermore, the representational approach of Shanks and Gluck's (1991) Consequential Region Model affords little freedom with regard either the complete internal development or ensuing learned adjustment of psychological space representations. The consequential region units activated by the presentation of a stimulus are completely determined by the psychological space representation of that stimulus, and the association of the consequential region units with regions of psychological space. Thus, the propagation of information from the input layer to the consequential region layer is both invariant and pre-determined. Effectively, a stimulus could be presented to the model solely in terms of its associated pattern of activation across the consequential region units.

---

<sup>5</sup> In unreported simulations (Kruschke, personal communication, February 1995), a gradient descent based learning rule for adjusting parameters in the radial basis function, as derived in Kruschke (1990), has been implemented within the ALEX model.



The model described by Shepard and Kannappan (1991) would appear to fare a little better in this regard. The layered representational structure employed allows the possibility of the principled adjustment of connections between stimulus points in psychological space and the consequential region units. Nevertheless, neither of the models adopting the consequential region approach provides a mechanism for the adjustment of the location of the consequential regions within psychological space. This weakness, which again constitutes a shortcoming since practical considerations regarding network resources limit the resolution and extent of the modelled psychological space, is potentially addressed by the learned adjustment of the exemplar and covering map units in, respectively, the ALEX and ALCOVE models.

Finally, the radial basis function approach appears to hold the most promise with regard to overcoming the most fundamental weakness of all of the models discussed above. This weakness concerns the models' inability to learn the psychological space representations they employ, and their subsequent inability to modify these representations in accordance with experience. The ALEX and ALCOVE models share with the models founded on the consequential region approach a reliance on the pre-determination of appropriate psychological space representation of the stimulus sets with which they are presented. The radial basis function approach, however, through its explicit modelling of the process whereby information represented at an input layer is transformed into a different internal representation, seems to offer a means of rectifying this shortcoming. The possibility exists of developing of a set of learning rules which modify the connection weights which perform these transformations, in such a way as to create internal psychological space representations.

---

### 3.3. The Connectionist Learning Of Psychological Space Representations

Guidance in the construction of such learning rules is assisted by noting that the pre-determination of stimulus representations in connectionist models employing the psychological space representational construct is primarily achieved through the use of the multidimensional scaling family of statistical techniques. Thus, a first step towards developing a connectionist model which is able to learn and modify psychological space internal representations is the development of a connectionist implementation of an appropriate multidimensional scaling algorithm. Hanson & Burr (1990) raise this possibility in acknowledging the "close analogy between multidimensional scaling and [connectionist] nets with hidden units" (p. 489), and further note that the error minimisation approach of multidimensional scaling (Kruskal 1964a, see also Shepard 1974) is particularly amenable to implementation using standard connectionist learning rules.

Smolensky (1988a) suggests that this task be approached directly and logically by developing "a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments" (p. 8). One attempt at precisely such a model is described by Rumelhart and

Todd (1993), and is slightly extended by Todd and Rumelhart (1995).

### 3.3.1. Rumelhart And Todd's Model

The basic architecture of the model employed by Rumelhart and Todd (1993) is shown in Figure 3.9. The model operates by generating a predicted similarity value following the pairwise presentation of stimuli from a stimulus set, then altering its internal representation of these stimuli in accordance with feedback provided concerning the 'correct' similarity value obtained from human subjects on an analogous task. The network architecture and connection weights which compare the internal representations of the two stimuli are both fixed, and adhere to the structural principle of multidimensional scaling that the derived similarity value varies monotonically with the similarity of the internal representations. Specifically, the comparison 'sub-network' or 'module' determines the featural similarity between the stimuli across each of their internally represented dimensions, before combining these dimension similarities into a final similarity measure. The learned mappings from the stimulus input layer to the internal representation layer are generated by backpropagation on the error imposed by feedback giving the 'correct' pairwise similarity value. These mappings are constrained to be identical for both of the 'modules' which perform this transformation - that is, the analogous connection weights between input and internal representation units for the first and second stimuli are always equal. In this way, the model learns a unique means of converting stimulus information into principled internal representations based on the psychological similarity of these stimuli.

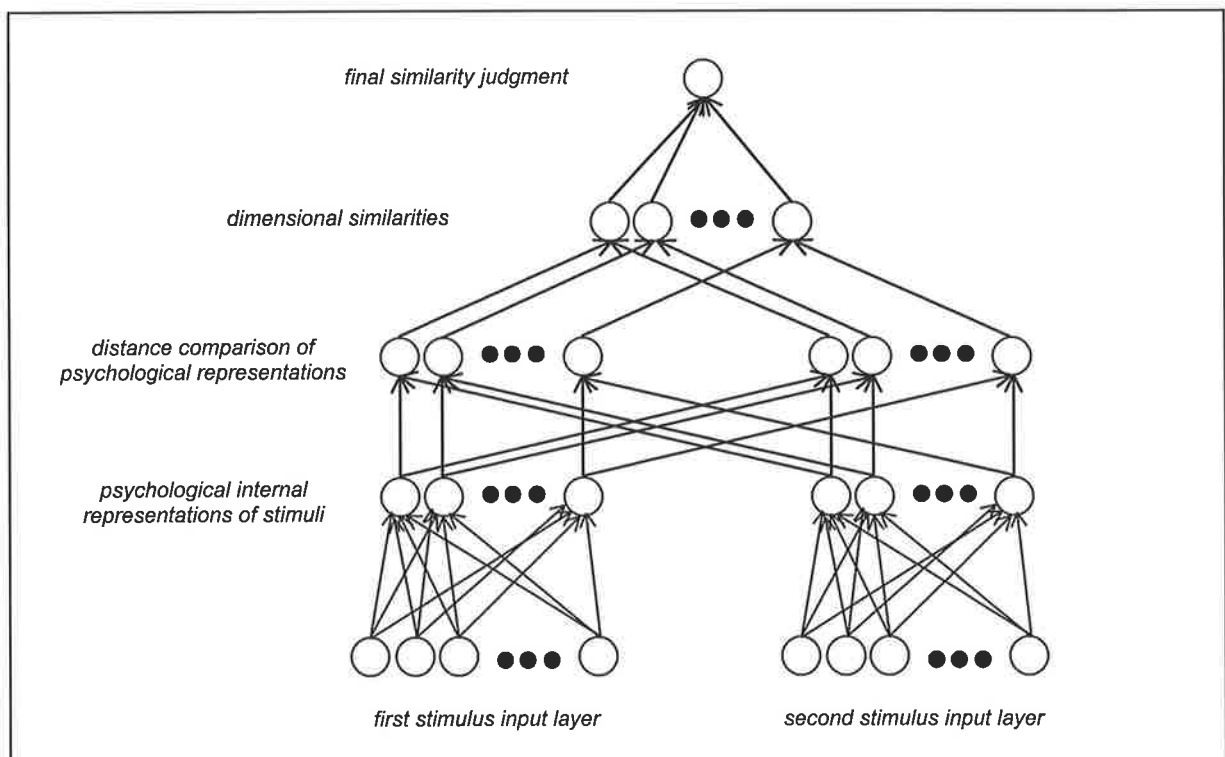


Figure 3.9. Rumelhart and Todd's (1993) connectionist implementation of multidimensional scaling. Adapted from Rumelhart and Todd (1993), Figures 1.15 and 1.17.

Rumelhart and Todd (1993) provide several demonstrations of the ability of their model to learn appropriate internal stimulus representations which resemble those derived from traditional multidimensional scaling techniques. Both local and distributed stimulus input representations are employed in these demonstrations, with the distributed representational approach accommodating an impressive generalisation ability within the model. After learning a mapping from input to internal representation for 31 members of a Morse code stimulus set, previously examined using traditional multidimensional scaling techniques (eg. Shepard 1980), the model exhibits generalisation capabilities by appropriately internally representing the 5 remaining stimuli without receiving feedback. In this, and all of these instantiations of the model, however, the dimensionality of the internal representational space is pre-determined.

Todd and Rumelhart (1995) explore various extensions and refinements of this basic modelling approach. The model is adapted to consider the 'feature matching' modelling of various cognitive phenomena founded on set theoretic manipulations of discrete mental representational structures (Tversky 1977). Todd and Rumelhart (1995) also discuss the natural way in which the model can learn to differentially weight stimulus dimensions, in response to both selective attention shifts of the type discussed above, and more permanent individual differences in dimensional salience, of the type usually revealed by INDSCAL (see Shepard 1980 for an overview) multidimensional scaling techniques. The limited ability of the model to mimic the 'stress' error minimisation of non-metric multidimensional scaling (Kruskal 1964a) is also explored, and the possibility of extending the model to perform the types of analyses identified with other clustering algorithms such as INCLUS and ADCLUS (again, see Shepard 1980) is also mentioned.

Todd and Rumelhart (1995) note that the inter-stimulus relation between similarity and distance generated by the model does not, beyond being monotonically decreasing, resemble a negative exponential function, in an apparent violation of the Universal Law of Generalization. As a remedy, the replacement of sigmoid activation functions with exponential decay activation functions for those units which calculate psychological similarity from featural differences is proposed. Subsequently, means by which both the City-block and Euclidean metrics could be accommodated by the model are also described although, unfortunately, the approach which is claimed to realise an Euclidean metric appears to correspond to the City-block metric, whilst the proposed City-block approach is difficult to reconcile with any familiar metric. More enlightening is Todd and Rumelhart's (1995) discussion of the means by which the model might be extended to self-determine the appropriate dimensionality of the psychological space in which its learned internal representations are embedded. In particular, the introduction of an additional term in the error function which serves to 'penalise' unnecessary internal representation units is suggested. Appropriately, the minimisation of such an augmented error by the model's learning rules would act to generate multidimensionally scaled internal representations in a psychological space of the

minimal possible dimensionality.

### 3.3.2. Connectionist Multidimensional Scaling In A Radial Basis Function Architecture

Despite the obvious merits of the connectionist multidimensional scaling modelling approach developed by Rumelhart and Todd (1993) and Todd and Rumelhart (1995), the previously articulated natural implementation of the psychological space representational construct and its associated generalisation functions and distance metrics by appropriately formulated radial basis function architectures maintains considerable suggestive force. In particular, the natural way in which the ALEX and ALCOVE models can be applied to the emulation of cognitive processes such as identification, recognition, categorisation, seems to contrast rather starkly with the contrived pairwise similarity measure production of Rumelhart and Todd's (1993) modelling approach. By focusing upon developing a canonical connectionist implementation of multidimensional scaling, the Rumelhart and Todd (1993) model sacrifices the possibility of ready extension to the modelling of broader human cognitive processes. In effect, Rumelhart and Todd's (1993) network fulfils the same role for cognitive connectionist modelling as traditional multidimensional scaling - that of the off-line derivation of psychological spaces which can be installed into other models.

As was argued in Chapter 1, an attempt to develop a connectionist model of the learning of human mental representational must conceive of conceptual structures as emergent phenomena, arising from the general cognitive action of humans in the world. Chalmers, French and Hofstadter (1991) suggest:

“In order to provide the kind of flexibility that is apparent in cognition, any fully cognitive model will probably require a continual interaction between the process of representation-building and the manipulation of those representations” (p. 8)

This argument that, contrary to the philosophy underpinning cognitive process models, mental representation and cognitive processing are inextricably intertwined is further developed in Chapter 6. There it is suggested that mental representations are appropriately conceived of as being reflections, as mediated by human cognitive processing, of the abstract physical principles which operate in the external world.

It would, therefore, seem highly desirable to develop a connectionist implementation of multidimensional scaling formulated within the radial basis function architecture utilised by the ALEX and ALCOVE models, with an ultimate goal of not only modelling the acquisition of psychological space internal representations, but of subsequently being able to accommodate within the same model the identification, categorisation, and other cognitive performance which is

underpinned by these newly acquired representations. Chapter 4 develops such a model.<sup>6</sup>

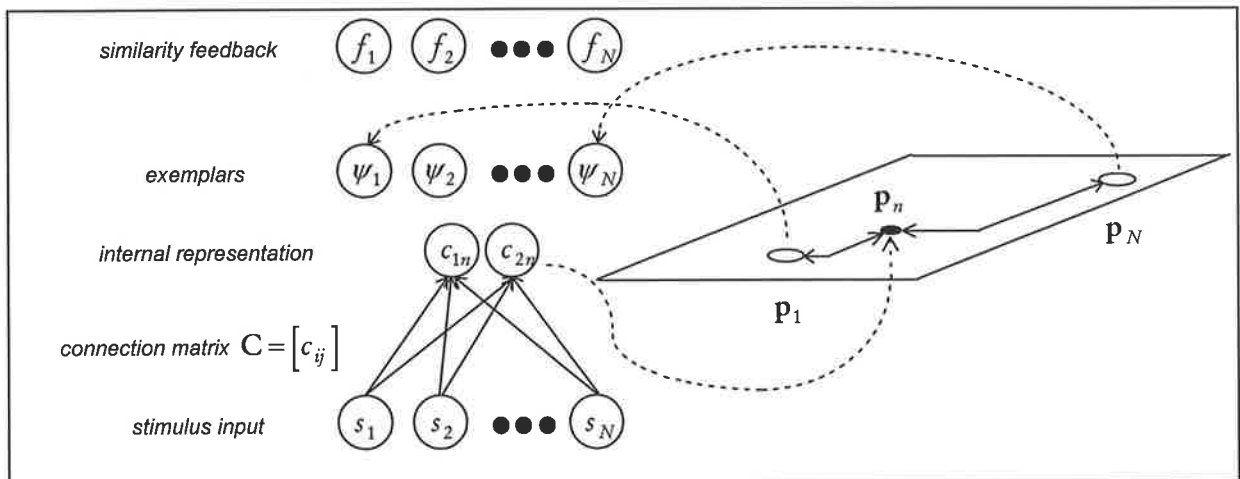
---

<sup>6</sup> Since the model described in Chapter 4 was developed, two further contributions to the connectionist multidimensional scaling literature have been reported. Webb (1995) presents a radial basis function implementation of multidimensional scaling which is entirely non-psychological in its construction and application, does not incorporate the notion of psychological space representation, and differs substantially from the model developed here in most aspects of its computational approach. Meanwhile, Bezdek and Pal (1995b) contribute analysis which further advances, without realising, the possibility of a self-organising map model learning multidimensionally scaled representational structures.

## Chapter 4: A Connectionist Multidimensional Scaling Model

This chapter describes and evaluates a connectionist model, formulated within a radial basis function architecture, which learns to represent internally a set of stimuli in terms of their multidimensionally scaled psychological space locations. The acceptance of the Universal Law of Generalization has strong implications for the development of such a model. Most fundamentally, the adoption of a radial basis function architecture becomes tenable because the metric nature of this architecture's internal representational space is justified. Effectively, the ability of non-metric multidimensional scaling techniques to recover metric information from ordinal data accommodates the assumption that psychological space is a coordinate metric space. As such, the model developed in this chapter constitutes a connectionist implementation of a metric multidimensional scaling algorithm.

The architecture of the model, and much of the nomenclature employed to describe the model, are displayed in Figure 4.1. The number of stimuli in the set which the model encounters is given by the number  $N$ , and  $N$  units are placed in the stimulus input, exemplar, and feedback layers. The units in each of these layers are created in one-to-one correspondence with the stimulus set, whilst each unit in the internal representation layer represents a single dimension of the internal representational space. In general, there are  $P$  such units in the internal representation layer, although Figure 4.1 assumes  $P = 2$  to assist in the graphical depiction of the representational space.



**Figure 4.1.** The architecture and nomenclature of the radial basis function connectionist implementation of multidimensional scaling.

The operation of the model is conveniently subdivided into processing and learning phases. Within the processing phase, the model generates a pattern of inter-stimulus similarity measures relating to a particular presented stimulus. Within the learning phase, externally derived information regarding appropriate values for these similarity measures is provided, and a learning rules modifies connection weights within the model to alter the internal representation of the presented stimulus.

---

## 4.1. Processing Phase

The processing phase consists of three main operations. First, a stimulus from the pre-determined stimulus set is presented to the model through the setting of activation values across the stimulus input layer. Secondly, the current internal representation of the presented stimulus is determined, using the connection weight matrix  $C$ , at the internal representation layer. Finally, the similarity between the internal representation of the presented stimulus, and the internal representations of all other members of the stimulus set, is calculated at the exemplar layer.

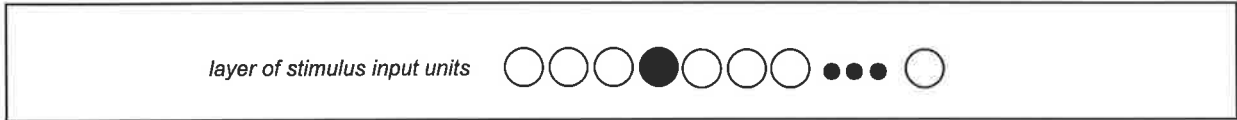
### 4.1.1. Stimulus Presentation

One of the most pervasive conclusions to be drawn from Section 2.1 concerns the self-delusionary dangers inherent in any utilisation of pre-abstraction in developing connectionist representations. Almost all ‘common-sense’, ‘logical’ or ‘intuitively reasonable’ stimulus representations, particularly ones which are not directly based on sensory description, contain some measure of psychological information. Clearly, it is inappropriate to provide the model with this type of information, since it constitutes a part of the mental representational structure which the model is supposed to learn.

Whilst, in some instances, it is probably true that this effect could be minimised, a model which is explicitly designed to model the acquisition of mental representation cannot afford to straddle the narrow divide between canonical stimulus description and solution-providing pre-abstraction. Thus, it would appear that the most expeditious approach to adopt, at least initially, is one in which the representation of stimuli contains absolutely no structural information. Fortunately, connectionist modelling can achieve this state of affairs if a model is required to learn the psychological internal representations of a finite number of stimuli: that is, if the stimuli can be described as a stimulus set.

The basis of such a representational scheme is depicted in Figure 4.2, which shows a stimulus input layer consisting of a number of units, created in one-to-one correspondence with the elements of the stimulus set. When a particular stimulus is presented to the model, the activation value of the input unit associated with that stimulus is made active, and all other units are made inactive. The rationale underlying this localist representational scheme is that connectionist modelling does not confer any sense of direction or other relation between unconnected units within a layer. That is, from the perspective of the internal operations of the model, there is no topological structure to distinguish a unit in the stimulus input layer from any other unit, despite the obvious relations that may be perceived from the graphical depiction of a linear array employed to describe the layer. As such, a representational scheme in which, following the presentation of a stimulus, one unit is active, and all others are inactive, conveys nothing more than nominal level information to the model. Effectively, the stimuli are presented as tokens, in

that their presence is unambiguously indicated, but no further information regarding their representational structure or relation to other stimuli is provided. Put another way, under this representational scheme, each stimulus representation is equally dissimilar (or similar) to every other representation, and hence the imposition of a similarity metric upon the stimulus input layer affords no representational insight.



*Figure 4.2. The localist input representation of the fourth member of a stimulus set.*

It should be conceded that this local approach to stimulus representation does impose some limitations on the model. Primarily, it requires that stimulus domains which are naturally described by a set of continuous parameters be quantised in some way. This restriction is particularly severe with regards sensory or physical stimulus domains of the type described in Section 2.2. For example, if the stimuli presented to a model take the form of vertical line segments of variable length, the necessity of replacing a description based on the value of a single (length) parameter with a potentially large discrete set of stimuli with different fixed lengths clearly does not constitute modelling parsimony.

Nonetheless, it is reasonable to assert that, at least initially, the advantages of the localist representational scheme outweigh the disadvantages. Psychology in general, and connectionist psychology in particular, has placed considerable emphasis on the representational structure of task domains - most notably members of various natural kinds - which are appropriately described as stimulus sets. In any case, the over-riding concern should be one of avoiding inappropriate pre-abstraction at almost any cost, and it is primarily on these grounds that this representational scheme is adopted for much of this thesis. In Chapter 9, however, consideration is given to ways in which the constraints of local representation can be overcome, and alternative schemes are used to extend the model towards addressing continuous stimulus domains.

For the moment, however, a particular stimulus is presented to the model by setting the corresponding unit in the stimulus input layer to an activated value of one, and deactivating all other stimulus input units with a value of zero. Formally, if stimulus  $n$  is presented ( $1 \leq n \leq N$ ), then the activation values of the stimulus input units,  $s_1, s_2, \dots, s_N$  are:

$$s_i = \begin{cases} 1 & \text{if } i = n \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

#### 4.1.2. Determining The Current Internal Representation

The pattern of activation across the stimulus input layer generates activation values at the



internal representation layer through the connection weight matrix  $C=[c_{ij}]$  with  $P$  rows and  $N$  columns. These activation values correspond to the coordinate values of the current internal representation of the presented stimulus. The  $i$ th such coordinate is given by:

$$\sum_{j=1}^N s_j c_{ij} = s_n c_{in} = c_{in} \quad (4.2)$$

where the simplification is a direct result of the adoption of local representation. Thus, the internal representation of the presented stimulus is given by  $\mathbf{p}_n = (c_{1n}, c_{2n}, \dots, c_{pn})$ .

#### 4.1.3. Calculating Psychological Similarity

Once the internal representational location of the stimulus has been established, the model activates the units in the exemplar layer. These units correspond to the elements of the stimulus set, and maintain a location within the internal representational space. Again, from the simplification evident in Equation 4.2, it is clear that the location of the  $j$ th exemplar unit within psychological space is given by  $\mathbf{p}_j = (c_{1j}, c_{2j}, \dots, c_{pj})$ . The activation value of this unit, denoted by  $\psi_j$ , measures the psychological similarity of the  $i$ th element of the stimulus set to the currently presented stimulus, and is calculated using the radial basis function linkage between the internal representation and exemplar layers. The complete specification of this linkage, however, requires both the adoption of a basis function and a decision regarding the distance metric operating within the representational space.

As was anticipated in Section 3.2.3, an acceptance of the Universal Law of Generalisation constrains the form of the basis function. In particular, the function relating psychological similarity to distance in psychological space should be an exponential decay function. Exponential decay functions, however, could sensibly be formulated to include at least three parameters, as follows:

$$f(D) = a + b \exp(-cD) \quad (4.3)$$

where  $f$  is the radial basis function determining psychological similarity,  $D$  is the distance in psychological space, and  $a$ ,  $b$ , and  $c$  are, respectively, translation, amplitude and decay parameters.

There are, however, compelling grounds on which each of these three parameters can be assumed to take particular fixed values. Following Shepard (1972, p. 73), it appears entirely reasonable to constrain the measures of psychological similarity arising from the application of this basis function to lie between values of zero, indicating complete dissimilarity, and one, indicating complete similarity. This assumption immediately fixes the values of  $a$  and  $b$  to be zero and one respectively.

The decay parameter,  $c$ , relates to what has variously been described as the “sensitivity”, (Nosofsky 1992, p. 33) or “specificity” (Kruschke 1992, p. 23) of psychological space, and effectively manipulates the spread of derived psychological similarity values. If  $c$  takes the value zero, for instance, all measures of psychological similarity, regardless of distance, will assume the value one. Conversely, as the value of  $c$  tends towards infinity, all psychological similarity measures approach zero. For intermediate  $c$  values, a broad spectrum of psychological similarity measures are derived. Clearly, however, manipulations of the distance variable,  $D$ , can achieve precisely the same effects. If all of the inter-stimulus distances in psychological space are increased by a large constant factor, the geometric equivalent of ‘spreading out’ all of the points in the space by a multiplication of the coordinate axes, and the value of  $c$  is held constant, the psychological similarity measures will again tend towards zero. Similarly, a contractive clustering of points will reduce distance values and generate psychological similarity values near one. Such rescalings are entirely permissible since the units of measurement within psychological spaces are essentially arbitrary. Thus, it seems clear that any variation in the value of  $c$  in Equation 4.3 can be counter-balanced by a distance manipulating rescaling of the axes of psychological space. Algebraically, this is reflected by the fact that an alteration of the value of either  $c$  or  $D$  will not alter the product  $cD$ , and hence will not alter the ultimate measure of psychological similarity,  $f(D)$ , providing the other value undergoes a compensatory modification. The implication of this relationship is that  $c$  can validly be fixed to any value greater than zero. For simplicity, the fixed value chosen is one.

The radial basis function assumed by the model is significantly simplified as a result of these considerations of appropriate parameter values for  $a$ ,  $b$ , and  $c$ , and becomes:

$$f(D) = \exp(-D) \tag{4.4}$$

Unfortunately, it is difficult to provide a similarly conclusive specification of the distance metric operating within the internal representational space. Primarily, these difficulties arise from the fact that no workable mechanism for the determination of the appropriate distance metric structure of the psychological space representation of a set of stimuli has been developed.

Often, the form of a distance metric employed within a psychological space is assumed on the basis of some knowledge or intuition regarding the nature of the stimuli involved (eg. Kruschke 1992, Nosofsky 1988b, Todd & Rumelhart 1995). If, however, more principled efforts are made to determine an appropriate form for the underlying metric, they usually take the form of the repeated application of a multidimensional scaling algorithm operating under various metrics (Kruskal 1964b, Shepard 1974, 1991). The final choice of distance metric then resides in observing which metric best accommodates the psychological space representation of the stimuli, in terms of minimising the error (or ‘stress’) measure. In some instances this may be a valid means of metric determination, in the sense of conducting a statistical analysis of psychological similarity data,

although it does appear susceptible to significant difficulties (see Arabie 1991, Borg 1982, Borg & Lingoes 1987, pp. 230-231). Specifically, the assumption that error values attained under the operation of different metrics can be compared directly is generally invalid, and this difficulty is enhanced by the suggestion that the likelihood of deriving sub-optimal representational solutions with artificially high error values also varies across different metrics.

In any case, the entire approach certainly does not constitute a reasonable model of a cognitive process. Unfortunately, a more realistic model of the way in which humans determine the appropriate metric basis for the mental representations they form has yet to be developed. Progress in this area probably requires theoretical advances in understanding the relationship between separable, integral and other types of stimuli and the developmental effects on these distinctions arising from processes such as the evolutionary internalisation of mental representation. Perhaps Shepard's (1987a) reconciliation of the Minkowskian family of distance metrics with the degree of correlation in the dimensional extension of consequential regions, as discussed in Section 3.1.3, could be viewed as the beginnings of the necessary theoretical development. Even more promising is Baxter's (1996) notion of 'canonical distance metrics' in representational spaces. This notion is founded upon the suggestion that the appropriateness of distance measures in a representational space may be determined from the functions which operate within that space. That is, the representations in a domain which are regarded as similar by the particular set of functions applied across that domain, should also be the representations which the distance metric regards as being similar. In this sense, the distance metric operating in a psychological space may well be the one which best reflects the cognitive functional transformations which are applied to a particular stimulus domain. Given the impressive quantitative detail of Baxter's (1996) development of these ideas, the construction of a workable mechanism for distance metric determination based on the notion of 'canonical distance metrics' should be a priority for future research.

Meanwhile, however, the construction of a model which learns psychological space representations is forced to assume the operation of one or other of the various possible Minkowski distance metrics by specifying, with reference to Equation 3.1, a value between 0 and 2 for the parameter  $r$ . Coupling a distance metric assumption made in this way with the basis function described by Equation 4.4, the generation of a pattern of inter-stimulus similarities across the exemplar layer is achieved by evaluating:

$$\psi_j = \exp(-\|\mathbf{p}_n, \mathbf{p}_i\|_r) \quad (4.5)$$

#### 4.1.4. Provision Of Feedback

The indices of psychological similarity generated across the exemplar layer effectively

constitute the output of the model. To the extent that these ‘predicted’ similarity values do not concur with values that are regarded as ‘correct’, whether empirically or otherwise generated, the current internal representation of the presented stimulus requires adjustment. The correct measures of psychological similarity are provided through the activation values,  $f_1, f_2, \dots, f_N$  of the similarity feedback layer.

---

## 4.2. Learning Phase

Following the provision of feedback regarding the correct pattern of inter-stimulus psychological similarities, a learning phase alters the internal representation of the presented stimulus through the operation of a learning rule which adopts a gradient descent approach to optimisation, and is derived from an error measure which encapsulates the representational principles of multidimensional scaling. This error measure is appropriately conceived of as the sum of two components, the first of which, called the similarity error, penalises the departure of the current representational structure from the required relationship between psychological similarity and distance in psychological space, and the second of which, called the dimensional error, penalises the use of surplus dimensions by the representational structure. The roles of the similarity and dimensional error components are essentially analogous to those of the ‘alpha’ and ‘beta’ forces, respectively, used in the original non-metric multidimensional algorithm described by Shepard (1962a).

### 4.2.1. Similarity Error

The similarity error,  $E^{sim}$ , is defined to be (proportional) to the sum of the squared difference between the predicted similarity values given across the exemplar layer, and the actual similarity values provided by the feedback layer, as follows:

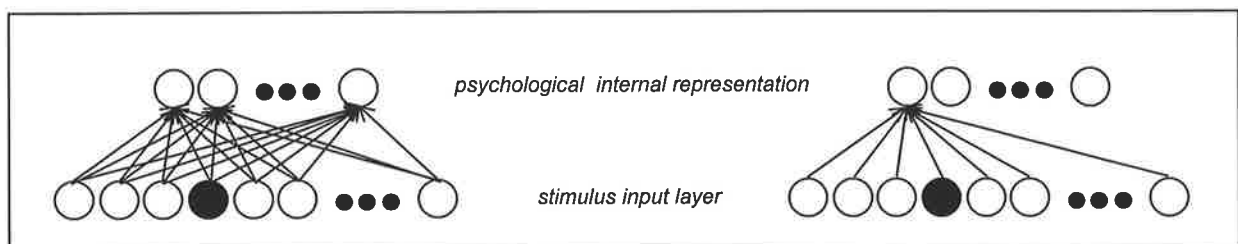
$$E^{sim} = \frac{1}{2} \sum_{j=1}^N (f_j - \psi_j)^2 \quad (4.6)$$

### 4.2.2. Dimensional Error

As noted in Section 3.3.1, connectionist analogues of the ‘beta’ forces employed for dimensionality reduction by Shepard’s (1962a) algorithm are most readily found in learning rules derived from error functions which contain penalty terms designed to optimise, in some sense, the structure of a network (see Ash & Cottrell 1995, Haykin 1994, pp. 207-209, Hertz, Krogh & Palmer 1991, p. 158, Reed & Marks 1995 for summaries). Given the correspondence between units in the internal representation layer and psychological space dimensions, the dimensional error measure employed by the model serves to reduce the dimensionality of psychological space by

defining a learning rule which modifies the connection weights in such a way as to allow the removal (or ‘pruning’) of units in this layer. A direct connectionist re-interpretation of ‘beta’ forces would involve an error measure which sought to maximise the variance of the set of distance measures in psychological space. Unfortunately, this approach seems constrained, through its reliance on a global variance measure, to the generation of learning rules which are significantly non-local in their operation, and thus negates one of the fundamental strengths of connectionist cognitive modelling identified in Chapter 1. The model developed here employs a different dimensional error measure, based on a ‘conjugate space’ analysis of the representational significance of the various dimensions of the current psychological space, which results in the specification of a learning rule requiring access to only those connection weights locally available from each internal representation unit.

This important difference in the requirements of the two approaches is graphically depicted in Figure 4.3. Measuring the variance of the entire set of inter-stimulus distances obviously requires access to the psychological space location of every stimulus, and hence, as is shown on the left of Figure 4.3, involves all of the weights connecting the stimulus input and internal representation layers. The approach formalised here, in contrast, adjusts each coordinate value of the location of the presented stimulus using only the so-called ‘instar’ (see Grossberg 1982) of connection weights arriving at the corresponding internal representation unit, as shown on the right of Figure 4.3. Admittedly, since the location of the current stimulus on each internal representational dimension is considered, every connection weight is at some point involved in the learning process. Nevertheless, the fact that a globally calculated variance measure is not required means that the connection weights need only be successively locally available. The key by which this more satisfactory state of affairs is attained is through developing an understanding of the relative representational importance of the various stimulus dimensions from the instar of connection weights.



*Figure 4.3. Locality of connection weight access required for dimensionality reduction when a stimulus, with stimulus node shown in black, is presented. The maximum variance approach, on the left, requires access to all weights whilst the proposed conjugate space method, on the right, requires access to only the instar of weights.*

The representational role of the individual stimulus dimensions in the internal representation layer can be gauged by considering the multidimensional coordinate space which is conjugate to the current psychological space. A conjugate space is constructed by associating elements of the stimulus set with the coordinate axes of the space, and representing the various stimulus

dimensions as points in this space, as defined by the values they assume with respect to each of the stimuli. As a simple example, consider a psychological space consisting of two three-dimensional points  $(a,b,c)$  and  $(d,e,f)$ . As is shown in Figure 4.4, the conjugate space consists of three two-dimensional points  $(a,d)$ ,  $(b,e)$  and  $(c,f)$ . In this conjugate space, the axes correspond to the two stimuli, whilst the points themselves represent the three original stimulus dimensions of the psychological space. In terms of the connection matrix  $C$ , the psychological space representation consists of the  $P$   $N$ -dimensional points defined by the  $P$   $N$ -dimensional vectors which are the rows of  $C$ , whilst the conjugate space consists of the  $N$   $P$ -dimensional points defined by the  $N$   $P$ -dimensional vectors which are the columns of  $C$ .

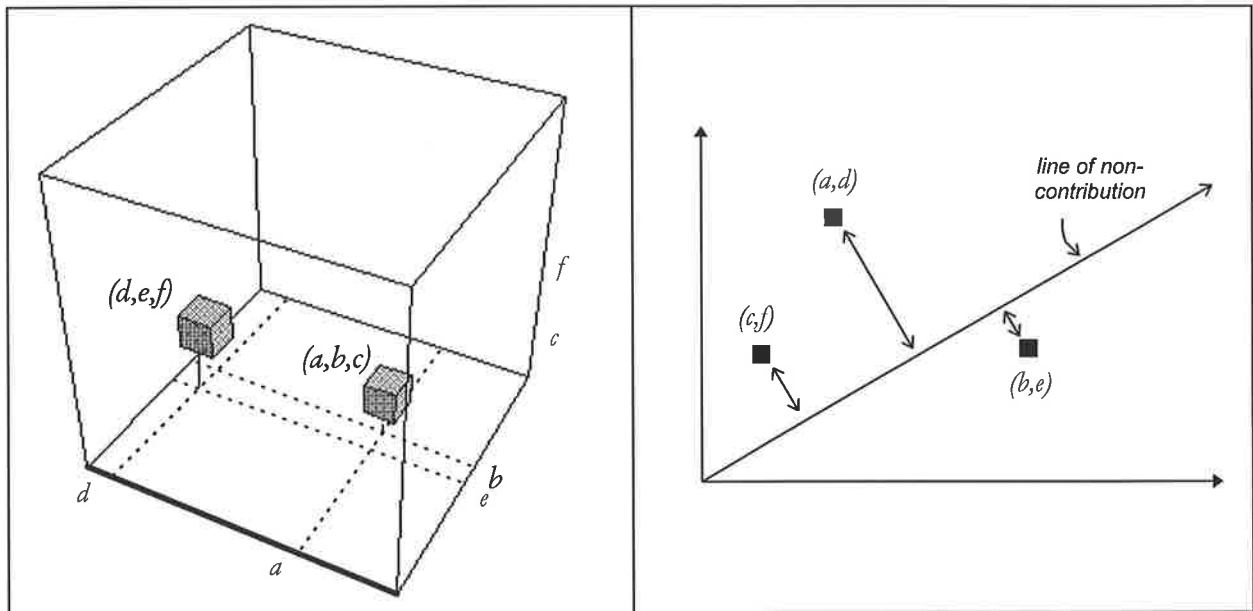


Figure 4.4. The relationship between psychological space, on the left, and its conjugate space with the line of non-contribution, on the right.

Effectively, the conjugate space reflects information regarding the representational contribution of the various stimulus dimensions of a psychological space. In particular, stimulus dimensions which assume relatively similar values for all stimuli are clearly making relatively less of a contribution towards accommodating the representational structure required by the similarity error measure. Geometrically, the line in an  $N$ -dimensional conjugate space, corresponding to an  $N$ -dimensional stimulus set, given by the parametric form:

$$(d_1 \ d_2 \ \dots \ d_N) = (t \ t \ \dots \ t), \quad t \text{ is a real number} \quad (4.7)$$

describes the set of points at which a particular stimulus dimension of psychological space is contributing no information to the representational structure. Consequently, this line will be termed the line of non-contribution, and any stimulus dimension coinciding with this line in conjugate space can be considered not to constitute a dimension of the relevant psychological space. In Figure 4.4, for example, the line of contribution is the familiar two-dimensional line described by the equation  $y = x$ , and the relative distances of the points  $(a,d)$ ,  $(b,e)$  and  $(c,f)$  from this line

indicate that the bold axis of the psychological space is making the most significant contribution to representational structure.

Furthermore, by defining the distance between a point and a line in the conjugate space to be proportional to the square of the usual Euclidean length of the perpendicular line segment from the point to the line involved, as indicated in Figure 4.4, the representational contribution of a stimulus dimension can reasonably be assumed to monotonically increase as this distance increases. Through using standard algebraic techniques (see, for example, Leithold 1986, p. 1033), this value can be calculated to give a measure of the representational importance of each stimulus dimension, which is denoted for the  $i$ th stimulus dimension by  $m_i$ , as follows:

$$m_i = \frac{1}{2} \sum_{j=1}^N (c_{ij} - \frac{1}{N} \sum_{k=1}^N c_{ik})^2 \quad (4.8)$$

The dimensional error term developed here is based on this measure of the relative contribution of the stimulus dimensions. In essence, the dimensional error seeks to attract stimulus dimensions which are near the line of non-contribution until they coincide and are therefore eliminated from the representational structure of psychological space. Clearly, the application of gradient descent minimisation techniques to any error measure which increases as the various stimulus dimensions become more distant from the line of non-contribution would constitute an instantiation of this approach. Beyond the requirement that the dimensional error measure results in the minimisation of the dimensionality of psychological space, however, it is also desirable that it interacts appropriately with the similarity error measure, in the sense that stable representational structures are readily derived.

To meet these ends, the dimensional error measure should produce attractive forces which decrease in magnitude very rapidly as the representational contribution measure increases. Loosely speaking, stimulus dimensions near the line of non-contribution should be subjected to strong attractive forces designed for their removal from the psychological space representation, whilst distant stimulus dimensions should remain essentially unaffected by these forces, thus being free to accommodate the representational structure required by the similarity error measure. The implication of these requirements for the form of the dimensional error measure is that it should monotonically increase as the representational contribution measure increases, but that the rate of this increase should monotonically decrease towards an asymptote of zero. These properties are partly evident in the form of the complexity penalty term introduced by Weigend, Rumelhart & Huberman (1991), which may recast in this context as:

$$Error = \frac{m^2}{a + m^2} \quad (4.9)$$

where  $a$  is a constant and  $m$  is a measure of representational contribution. The behaviour of this error measure, and its first derivative, are shown in Figure 4.5. Whilst the error measure clearly increases as  $m$  increases and also tends towards a gradient of zero, the behaviour of the gradient for small values of  $m$  is inappropriate. Specifically, the decrease in the gradient as the representation error measure approaches zero may prevent stimulus dimensions near the line of non-contribution in conjugate space from being subjected to the necessary strong attractive forces required for their removal from psychological space.

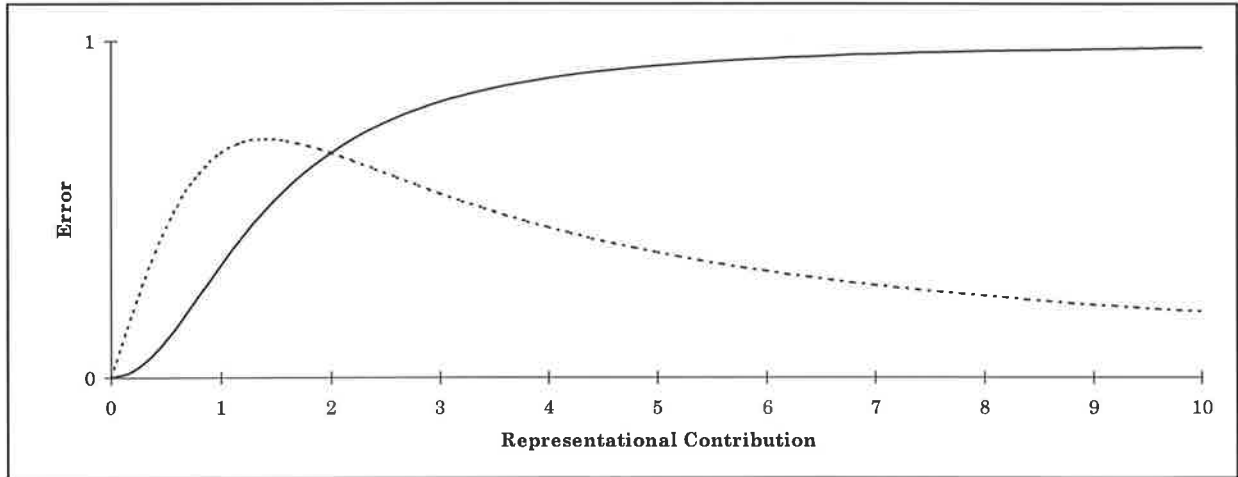


Figure 4.5. The error measure described by Equation 4.9, with  $a=2$  (solid line), and its first derivative (dotted line).

The dimensional error measure used in this model, therefore, employs a negative exponential decay function of the form:

$$Error = -\alpha \exp(-\beta m) + \alpha \quad (4.10)$$

with parameters  $\alpha$  and  $\beta$ , on the grounds that, as is evident from Figure 4.6, it completely satisfies the dual requirements of exhibiting a monotone increase in error, and a monotone decrease in the rate of change of this error, as the representational contribution measure increases. The function given in Equation 4.10 has the advantage of possessing a manipulable range of the interval  $(0, \alpha]$  across the relevant domain of non-negative real numbers, and is continuous, differentiable and analytically easy to manipulate. Consequently, the dimensional error measure for the  $i$ th stimulus dimension, denoted by  $E_i^{dim}$ , takes the form:

$$E_i^{dim} = -\frac{1}{\beta \lambda_c} \exp(-\beta m_i) + \frac{1}{\beta \lambda_c} \quad (4.11)$$

where the value of the  $\alpha$  parameter, as discussed in Section 4.2.3, has been chosen to simplify the



derivation and interpretation of the learning rule.

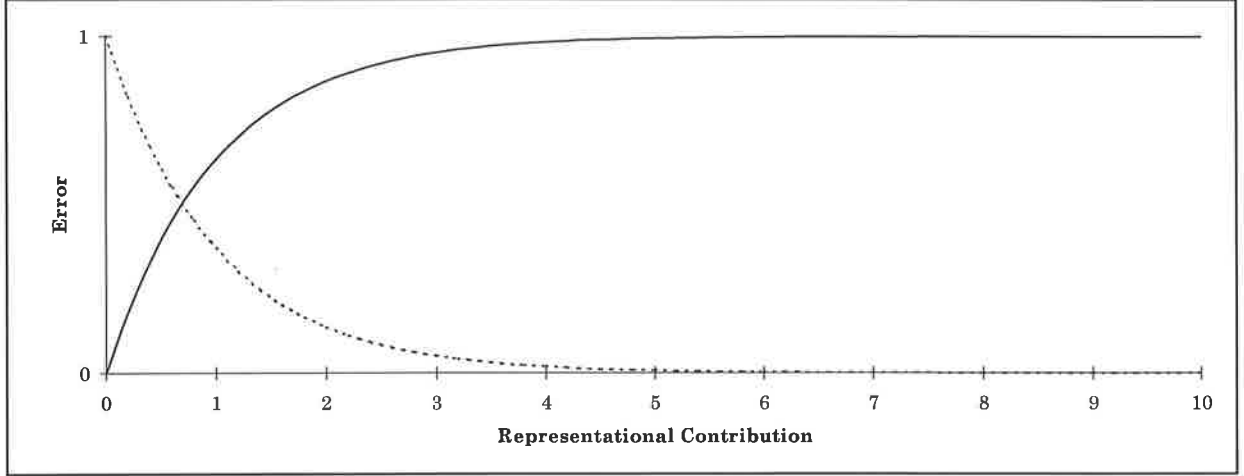


Figure 4.6. The error measure described by Equation 4.10, with  $\alpha=\beta=1$  (solid line), and its first derivative (dotted line).

Finally, the total dimensional error measure across all of the stimulus dimensions, denoted by  $E^{dim}$ , can now be defined as:

$$E^{dim} = \sum_{i=1}^P E_i^{dim} \quad (4.12)$$

#### 4.2.3. Derivation Of The Learning Rule

The learning rule employed by the model is derived from a total error measure,  $E^{tot}$ , given simply by:

$$E^{tot} = E^{sim} + E^{dim} \quad (4.13)$$

and acts to change the  $P$  connection weights in the  $n$ th column of  $C$ , which define the internal representation of the currently presented stimulus. The gradient descent optimisation principle adopted by the model results in a learning rule of the form:

$$c_{in}^{new} = c_{in}^{old} - \lambda_c \frac{\partial E^{tot}}{\partial c_{in}} \quad (4.14)$$

where  $\lambda_c$  is the learning rate parameter anticipated in Equation 4.11. The calculation of the required partial derivative is achieved by first observing:

$$\begin{aligned} \frac{\partial E^{tot}}{\partial \hat{\alpha}_{in}} &= \frac{\partial}{\partial \hat{\alpha}_{in}} (E^{sim} + E^{dim}) \\ &= \frac{\partial}{\partial \hat{\alpha}_{in}} (E^{sim} + E_i^{dim}) \\ &= \frac{\partial E^{sim}}{\partial \hat{\alpha}_{in}} + \frac{\partial E_i^{dim}}{\partial \hat{\alpha}_{in}} \end{aligned} \quad (4.15)$$

The partial derivative of the similarity is found as follows:

$$\begin{aligned}
\frac{\partial \mathcal{E}^{sim}}{\partial \alpha_{in}} &= \frac{\partial}{\partial \alpha_{in}} \frac{1}{2} \sum_{j=1}^N (f_j - \psi_j)^2 & (4.16) \\
&= -\sum_{j=1}^N (f_j - \psi_j) \frac{\partial \psi_j}{\partial \alpha_{in}} \\
&= -\sum_{j=1}^N (f_j - \psi_j) \frac{\partial}{\partial \alpha_{in}} \exp(-\|\mathbf{p}_n, \mathbf{p}_j\|_r) \\
&= \sum_{j=1}^N (f_j - \psi_j) \exp(-\|\mathbf{p}_n, \mathbf{p}_j\|_r) \frac{\partial \|\mathbf{p}_n, \mathbf{p}_j\|_r}{\partial \alpha_{in}} \\
&= \sum_{j=1}^N (f_j - \psi_j) \psi_j \frac{\partial \|\mathbf{p}_n, \mathbf{p}_j\|_r}{\partial \alpha_{in}} \\
&= \sum_{j=1}^N (f_j - \psi_j) \psi_j \frac{\partial}{\partial \alpha_{in}} \left( \sum_{k=1}^P |c_{kn} - c_{kj}|^r \right)^{\frac{1}{r}} \\
&= \sum_{j=1}^N (f_j - \psi_j) \psi_j \frac{1}{r} \left( \sum_{k=1}^P |c_{kn} - c_{kj}|^r \right)^{\frac{1-r}{r}} \frac{\partial}{\partial \alpha_{in}} |c_{in} - c_{ij}|^r \\
&= \sum_{j=1}^N (f_j - \psi_j) \psi_j \frac{1}{r} \|\mathbf{p}_n, \mathbf{p}_j\|_r^{1-r} \frac{\partial}{\partial \alpha_{in}} |c_{in} - c_{ij}|^r \\
&= \sum_{j=1}^N (f_j - \psi_j) \psi_j \|\mathbf{p}_n, \mathbf{p}_j\|_r^{1-r} |c_{in} - c_{ij}|^{r-1} \text{sgn}(c_{in} - c_{ij})
\end{aligned}$$

where  $\text{sgn}(\cdot)$  is the signum operator which takes the value -1 for a negative argument, +1 for a positive argument, and is 0 otherwise. The partial derivative of the dimensional error, through ignoring the effect a change on the given stimulus dimension for the presented stimulus has on the average taken across all stimuli, is appropriately approximated by:

$$\begin{aligned}
\frac{\partial \mathcal{E}_i^{dim}}{\partial \alpha_{in}} &= \frac{\partial}{\partial \alpha_{in}} \left( -\frac{1}{\beta \lambda_c} \exp(-\beta m_i) + \frac{1}{\beta \lambda_c} \right) & (4.17) \\
&= \frac{1}{\lambda_c} \exp(-\beta m_i) \frac{\partial m_i}{\partial \alpha_{in}} \\
&= \frac{1}{\lambda_c} \exp(-\beta m_i) \frac{\partial}{\partial \alpha_{in}} \frac{1}{2} \sum_{j=1}^N \left( c_{ij} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right)^2 \\
&\approx \frac{1}{\lambda_c} \exp(-\beta m_i) \frac{\partial}{\partial \alpha_{in}} \frac{1}{2} \left( c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right)^2 \\
&\approx \frac{1}{\lambda_c} \exp(-\beta m_i) \left( c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right)
\end{aligned}$$

It is the first of these approximations which allows the learning rule to act locally, as discussed in Section 4.2.2.

Substituting Equations 4.16 and 4.17 into Equation 4.15 defines the required partial

derivative, which, upon further substitution into Equation 4.14, results in the learning rule:

$$\begin{aligned}
c_{in}^{new} &= c_{in}^{old} - \lambda_c \frac{\partial E^{tot}}{\partial c_{in}} \\
&= c_{in}^{old} - \lambda_c \left( \sum_{j=1}^N (f_j - \psi_j) \psi_j \|\mathbf{P}_n, \mathbf{P}_j\|_r^{1-r} |c_{in} - c_{ij}|^{r-1} \text{sgn}(c_{in} - c_{ij}) \right) \\
&\quad + \frac{1}{\lambda_c} \exp(-\beta m_i) \left( c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right) \\
&= c_{in}^{old} - \lambda_c \sum_{j=1}^N (f_j - \psi_j) \psi_j \|\mathbf{P}_n, \mathbf{P}_j\|_r^{1-r} |c_{in} - c_{ij}|^{r-1} \text{sgn}(c_{in} - c_{ij}) \\
&\quad - \exp(-\beta m_i) \left( c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right)
\end{aligned} \tag{4.18}$$

The precise form of the learning rule is dependent upon the specification of a value for the parameter  $r$ , in accordance with distance metric assumed to operate within psychological space. Of particular interest are the values  $r = 1$  and  $r = 2$ , corresponding to separable and integral stimulus sets. The specific learning rules for these two cases may be evaluated through substituting the appropriate  $r$  value into the partial derivative of the similarity error given in Equation 4.16, since the partial derivative of the dimensional error is fixed over all possible distance metrics.

When  $r = 1$ , the partial derivative of the similarity error becomes:

$$\left. \frac{\partial E^{sim}}{\partial c_{in}} \right|_{r=1} = \sum_{j=1}^N (f_j - \psi_j) \psi_j \text{sgn}(c_{in} - c_{ij}) \tag{4.19}$$

giving the separable stimulus learning rule:

$$c_{in}^{new} = c_{in}^{old} - \lambda_c \sum_{j=1}^N (f_j - \psi_j) \psi_j \text{sgn}(c_{in} - c_{ij}) - \exp(-\beta m_i) \left( c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik} \right) \tag{4.20}$$

When  $r = 2$ , the partial derivative of the similarity error is given by:

$$\begin{aligned}
\left. \frac{\partial E^{sim}}{\partial c_{in}} \right|_{r=2} &= \sum_{j=1}^N (f_j - \psi_j) \psi_j \|\mathbf{P}_n, \mathbf{P}_j\|_2^{-1} |c_{in} - c_{ij}| \text{sgn}(c_{in} - c_{ij}) \\
&= \sum_{j=1}^N \frac{(f_j - \psi_j) \psi_j}{\|\mathbf{P}_n, \mathbf{P}_j\|_2} |c_{in} - c_{ij}| \text{sgn}(c_{in} - c_{ij}) \\
&= \sum_{j=1}^N \frac{(f_j - \psi_j) \psi_j (c_{in} - c_{ij})}{\|\mathbf{P}_n, \mathbf{P}_j\|_2}
\end{aligned} \tag{4.21}$$

resulting in the integral stimulus learning rule:

$$c_{in}^{new} = c_{in}^{old} - \lambda_c \sum_{j=1}^N \frac{(f_j - \psi_j) \psi_j (c_{in} - c_{ij})}{\|\mathbf{p}_n, \mathbf{p}_j\|_2} - \exp(-\beta m_i) (c_{in} - \frac{1}{N} \sum_{k=1}^N c_{in}) \quad (4.22)$$

---

### 4.3. Construction And Interpretation Of The Model

Given a particular set of stimuli, and known inter-stimulus similarity relationships, a model is constructed by establishing stimulus input, exemplar, and feedback layers which contain units in one-to-one correspondence with the elements of the stimulus set. The internal representation layer is constructed so as to contain enough units to overestimate the appropriate dimensionality of the psychological space representation of the stimulus set. It seems reasonable to assert that there is an upper bound on the dimensionality of the psychological spaces employed by humans, although the definitive quantification of this bound appears problematic. Perhaps some guidance might be sought in estimates of short term memory capacity (eg. Miller 1956). For the moment, however, it should be noted that the psychological spaces employed in this thesis resemble the vast majority of those discussed in the literature in the sense that they have low dimensionality. Therefore, it does not seem inappropriate to somewhat arbitrarily construct models with, say, six internal representation units, on the grounds that the psychological spaces the model must learn contain significantly fewer than six dimensions.

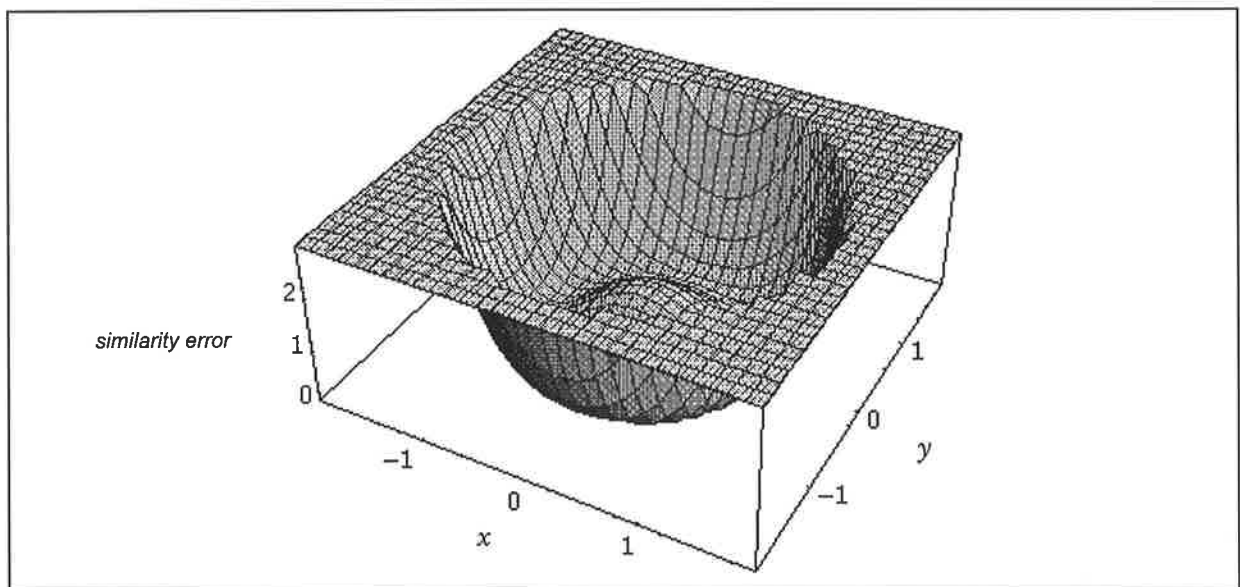
The psychological space representation developed by the model stabilises when the tendency towards dimensionality reduction is counterbalanced by the requirement to achieve the specified inter-stimulus patterns of similarity. At this point, the measures of representational contribution for the various stimulus dimensions are readily divided into two classes. One set of stimulus dimensions will have dimensional error values near zero. The remaining stimulus dimensions will have significantly larger dimensional error values, and are appropriately considered to be the axes of the derived psychological space representational structure.

It is the choice of  $\alpha$  in Equation 4.10, when coupled with the general form of the dimensional error measure, which gives rise to this division between included and removed stimulus dimensions, as an analysis of the final term of the learning rules given in Equation 4.18 reveals. This final term arises solely from the differentiation of the dimensional error component of the total error measure, and is appropriately interpreted through an examination of its two factors. The second factor acts to move the currently considered stimulus dimension of the presented stimulus towards the line of non-contribution at which the stimulus dimension can be removed from the derived psychological space representation. The first factor, therefore, acts as an adaptive learning rate parameter which varies according to the measure of representational contribution. Stimulus dimensions near the line of non-contribution will quickly converge towards

the line, resulting in their removal from the psychological space representation. In contrast, stimulus dimensions with large measures of representational contribution correspond to near-zero learning rate parameters, and will remain essentially unaffected by the dimensional error measure.

Graphically, this state of affairs is evident from an examination of the construction of the total error measure in terms of its similarity and dimensional components. Consider, for example, a two dimensional internal representation layer in which two stimulus points are to be located such that they are one unit apart. Without loss of generality, one of these points may be fixed at the origin, and the second point denoted by the coordinates  $(x,y)$ . Although the unit 'circle' (the form of which depends upon the distance metric operating within the space) consists of an infinite number of points on which  $(x,y)$  could be located to satisfy the inter-point distance constraint, the requirement of minimal dimensionality implies that the four points  $(1,0)$ ,  $(0,1)$ ,  $(-1,0)$ ,  $(0,-1)$  constitute the appropriate representational solution.

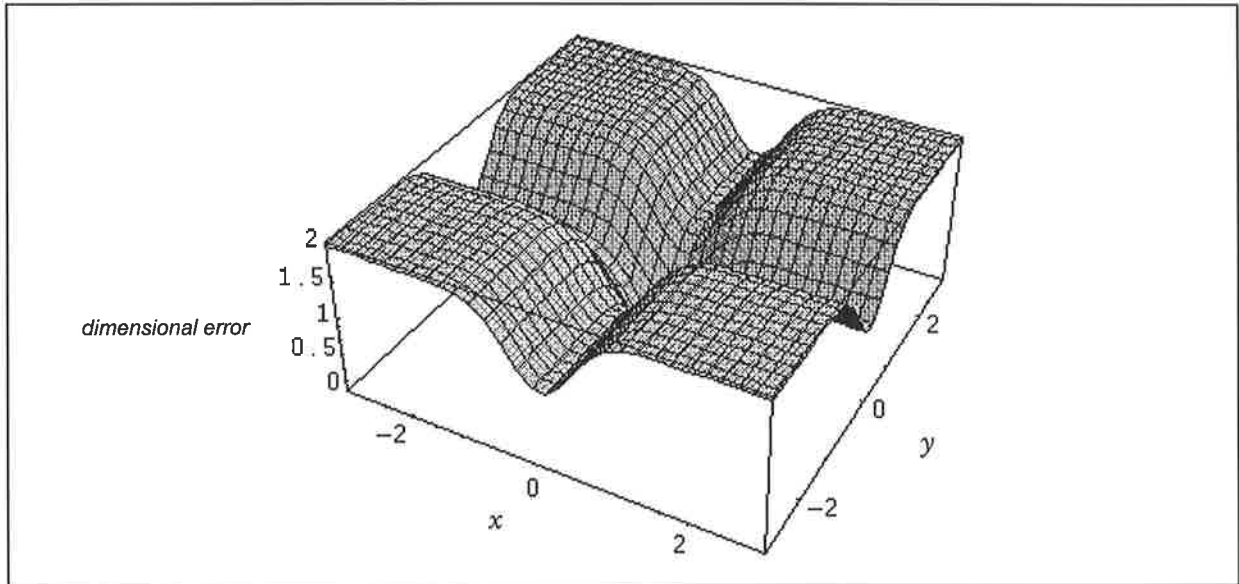
Assuming the Euclidean distance metric, the similarity error surface as the location of the second point varies is shown in Figure 4.7, and contains a circular 'trough' one unit from the origin which corresponds to a minimum error. It is this trough which consists of all points satisfying the inter-point distance constraint, as discussed above.



*Figure 4.7. The form of the similarity error measure across a two dimensional internal representational layer for the two point problem. The surface has been thresholded at an error value of 3 to assist graphical depiction.*

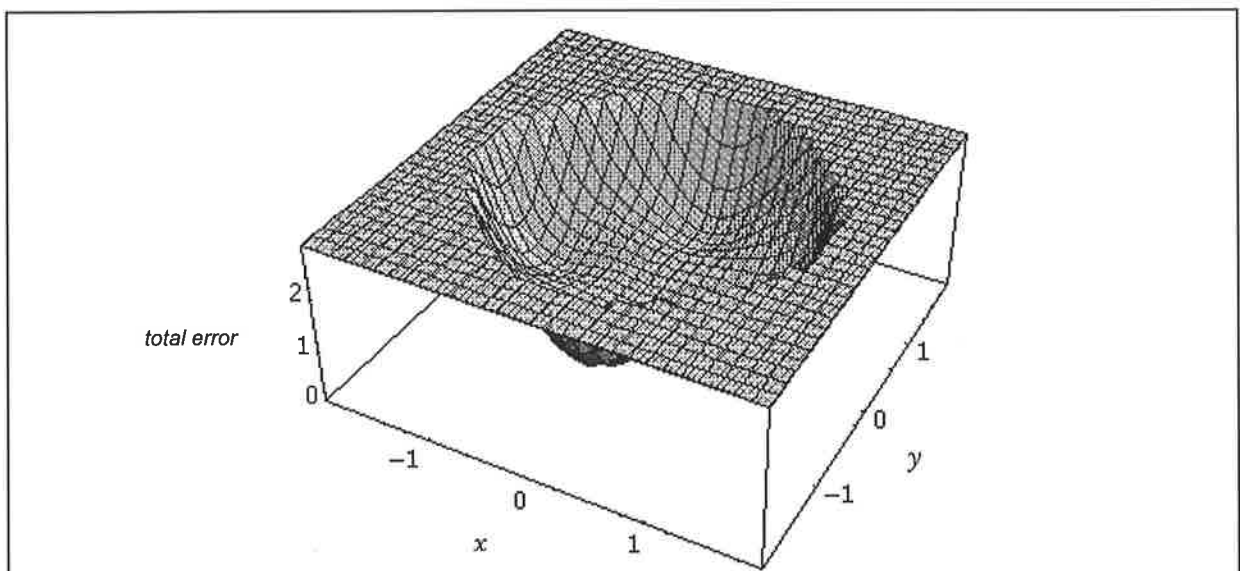
Since one point is fixed at the origin, the measure of representational contribution reduces to a measure of the distance of the point  $(x,y)$  from each of the coordinate axes. Thus, the dimensional error surface generated as the location of the second point varies is of the general form shown in Figure 4.8. Recalling that the learning rule performs gradient-descent on this surface, the large 'plateau' across the region in which either or both stimulus dimensions are making a significant representational contribution implies that the dimensional error has negligible impact on the alterations made by the learning rule in this case. The deep 'valleys' along the coordinate axes,

however, indicate that, when a stimulus dimension is making a near-zero contribution, the application of gradient-descent principles through the learning rule will result in the rapid removal of that stimulus dimension.



*Figure 4.8. The form of the dimensional error measure across a two dimensional internal representational layer for the two point problem.*

The construction of the total error through the summation of the similarity and dimensional errors for this two point problem is shown in Figure 4.9. It can be seen that the basins of attraction under a gradient-descent scheme correspond to the appropriate solutions points  $(1,0)$ ,  $(0,1)$ ,  $(-1,0)$ ,  $(0,-1)$  as described earlier. In addition, the negligible influence of the dimensional error when both stimulus dimensions are representing information is again evident by noting that the total error surface has the same form as the similarity error surface in those regions of the two dimensional domain in which both  $x$  and  $y$  are not near their respective coordinate axes. In this sense, the way in which the total error is defined can be seen to seek the accommodation of the similarity relations between stimulus points before attempting to remove surplus stimulus dimensions.



*Figure 4.9. The form of the total error measure across a two dimensional internal representational layer for the two point problem. The surface has been thresholded at an error value of 3 to assist graphical depiction.*

## Chapter 5: Evaluation Of The Connectionist Multidimensional Scaling Model

This chapter presents several demonstrations of the model developed in Chapter 4, involving both psychologically separable and integral stimulus sets, then considers the model's sensitivity to the values of its various parameters, and finally evaluates the model's ability to avoid the generation of representational configurations which are only locally optimal.

### 5.1. Demonstrations Of The Model

#### 5.1.1. Colour Model

To demonstrate the model's ability to learn the psychological space internal representation of integral stimuli, data reported by Ekman (1954, Table 1) was used. These data are replicated in Table 5.1, and gives the judged similarities between 14 colours with the wavelengths (measured in nanometres) specified.

Table 5.1. Colour similarity matrix

	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	1.000	0.860	0.420	0.420	0.180	0.060	0.070	0.040	0.020	0.070	0.090	0.120	0.130	0.160
445	0.860	1.000	0.500	0.440	0.220	0.090	0.070	0.070	0.020	0.040	0.070	0.110	0.130	0.140
465	0.420	0.500	1.000	0.810	0.470	0.170	0.100	0.080	0.020	0.010	0.020	0.010	0.050	0.030
472	0.420	0.440	0.810	1.000	0.540	0.250	0.100	0.090	0.020	0.010	0.000	0.010	0.020	0.040
490	0.180	0.220	0.470	0.540	1.000	0.610	0.310	0.260	0.070	0.020	0.020	0.010	0.020	0.000
504	0.060	0.090	0.170	0.250	0.610	1.000	0.620	0.450	0.140	0.080	0.020	0.020	0.020	0.010
537	0.070	0.070	0.100	0.100	0.310	0.620	1.000	0.730	0.220	0.140	0.050	0.020	0.020	0.000
555	0.040	0.070	0.080	0.090	0.260	0.450	0.730	1.000	0.330	0.190	0.040	0.030	0.020	0.020
584	0.020	0.020	0.020	0.020	0.070	0.140	0.220	0.330	1.000	0.580	0.370	0.270	0.200	0.230
600	0.070	0.040	0.010	0.010	0.020	0.080	0.140	0.190	0.580	1.000	0.740	0.500	0.410	0.280
610	0.090	0.070	0.020	0.000	0.020	0.020	0.050	0.040	0.370	0.740	1.000	0.760	0.620	0.550
628	0.120	0.110	0.010	0.010	0.010	0.020	0.020	0.030	0.270	0.500	0.760	1.000	0.850	0.680
651	0.130	0.130	0.050	0.020	0.020	0.020	0.020	0.020	0.200	0.410	0.620	0.850	1.000	0.760
674	0.160	0.140	0.030	0.040	0.000	0.010	0.000	0.020	0.230	0.280	0.550	0.680	0.760	1.000

The colour model consisted of 14 units in the stimulus input, exemplar and feedback layers, and 6 units in the internal representation layer. The initial location of each stimulus in the internal representational space was generated by selecting a random number in the range (0,0.5) independently for each stimulus on each representational dimension. The learning rate parameter  $\lambda_c$  and dimensionality reduction parameter  $\beta$  were set to values of 0.1 and 10 respectively.

On each of 1,000 trials, a randomly selected stimulus was presented to the model, causing the generation of predicted similarity values across the entire stimulus set. The appropriate row (or, equivalently, column, since the matrix is symmetric) of inter-stimulus similarities from Table 5.1 was then provided in the feedback layer, and the Euclidean learning rule was applied to modify the internal representation of the presented stimulus.

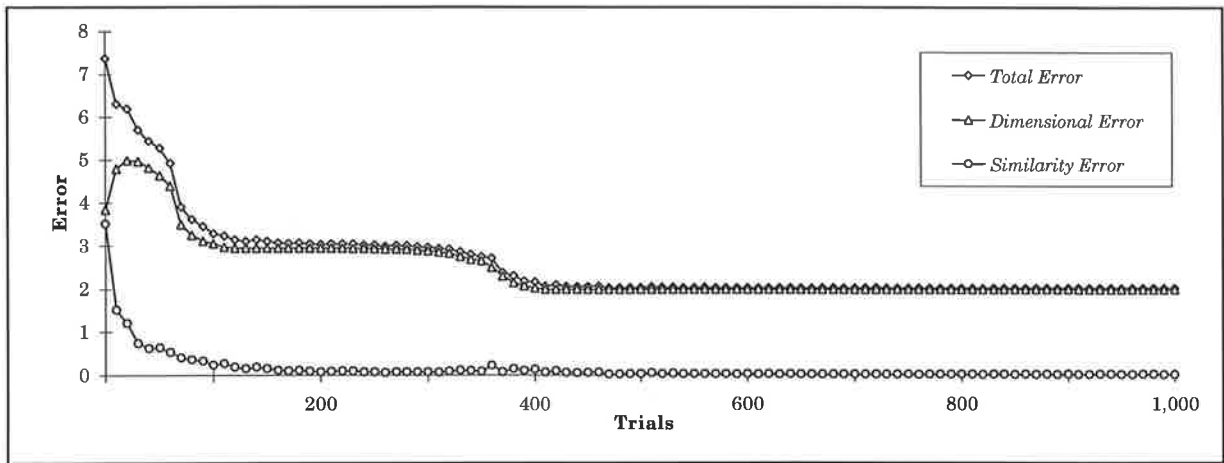


Figure 5.1. The pattern of change of the three error measures across 1,000 trials for the colour model.

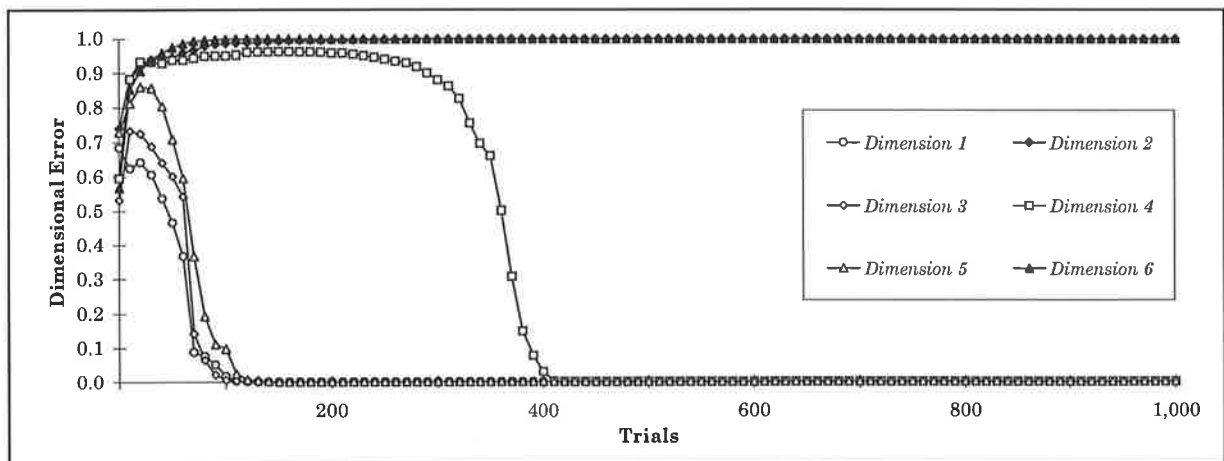


Figure 5.2. The breakdown of dimensional error across the 6 component stimulus dimensions for the colour model.

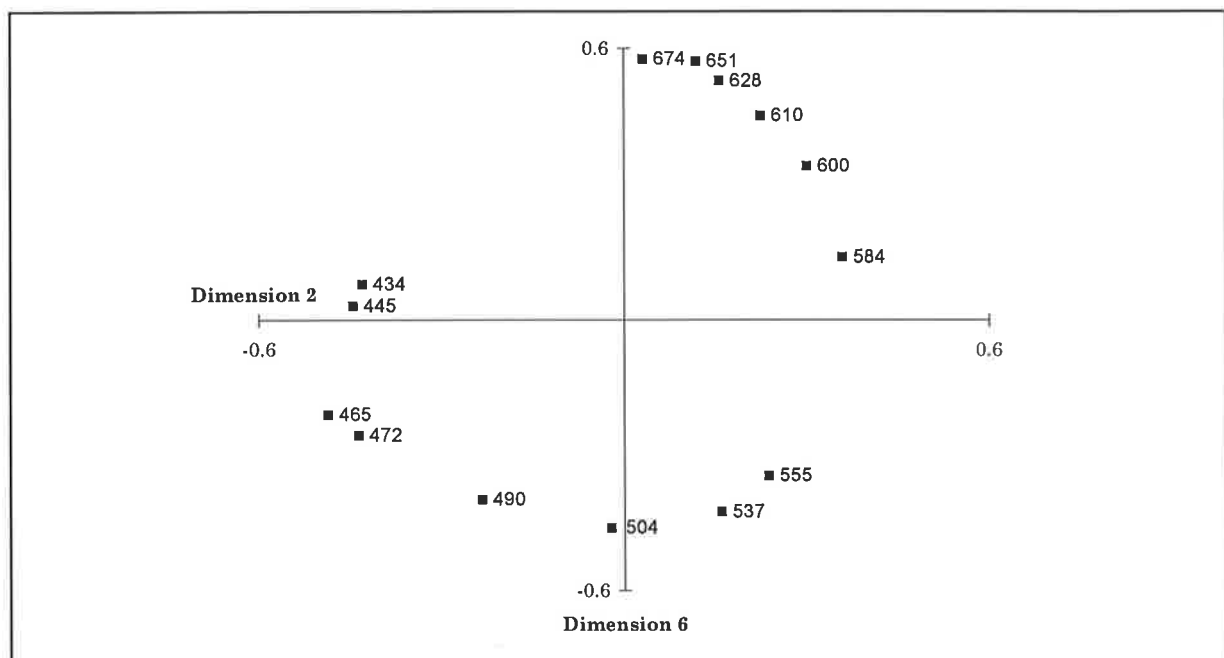


Figure 5.3. The final two-dimensional internal representation developed by the colour model.

The pattern of change of the three error measures - total error, dimensional error, and similarity error - across the 1,000 trials is shown in Figure 5.1. The similarity error decreases rapidly to a negligible value which it maintains. The dimensional error initially increases as the



internal representations of the stimuli are altered to satisfy the representational dictates of the similarity error, but then decreases to a final stable value of 2.

It is most instructive to decompose the dimensional error measure across the six stimulus dimensions (recall Equation 4.12), as shown in Figure 5.2. Three of the stimulus dimensions, corresponding to first, third and fifth unit in the internal representation layer, quickly attain negligible dimensional error values and hence are removed from the derived representational structure. A fourth stimulus dimension is removed after about 400 trials, resulting in a learned internal representation of the stimulus set which consists only of stimulus dimensions two and six. This two-dimensional representation is shown in Figure 5.3, and closely accords with the ‘colour circle’ representational structure previously revealed through the application of traditional multidimensional scaling algorithms to the same data (eg. Shepard 1962b).

### 5.1.2. Flower Pot Model

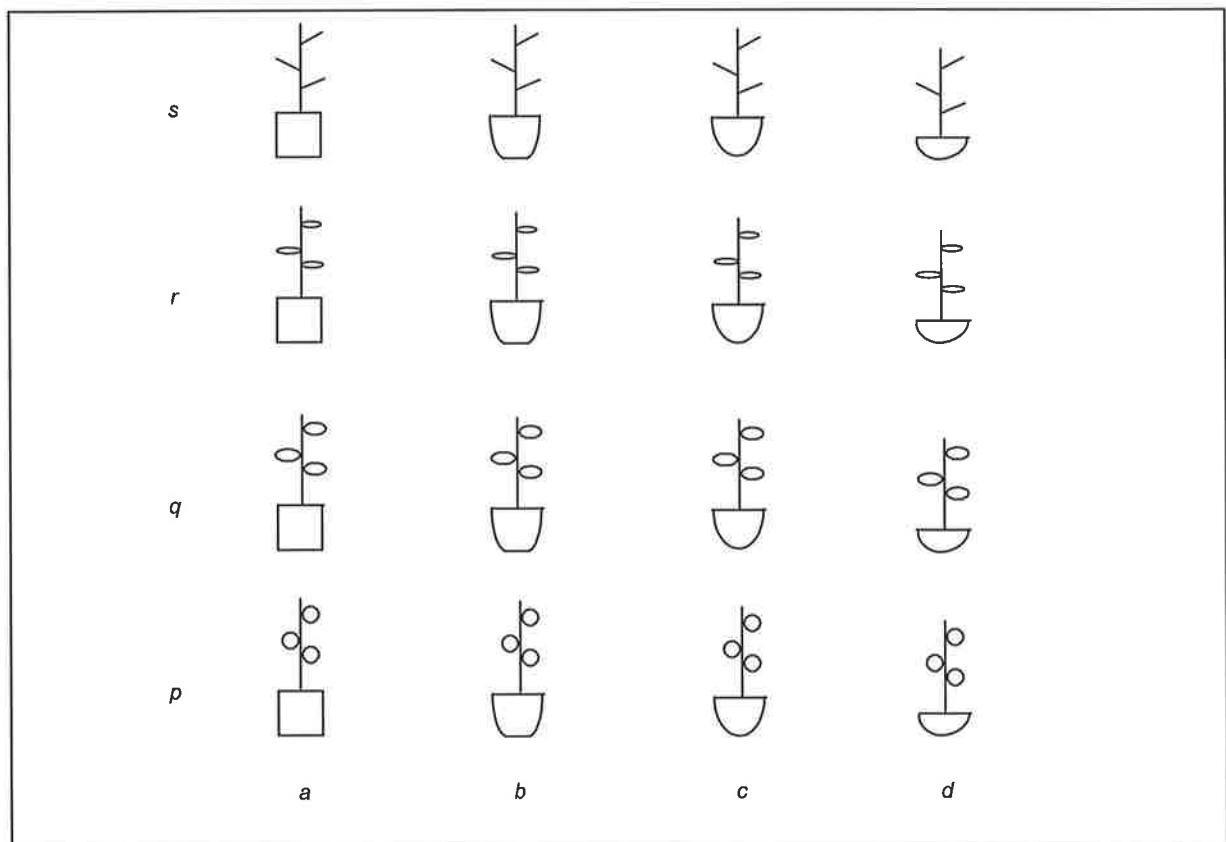


Figure 5.4. The 16 flower pot stimuli. Based on Gati and Tversky (1982), Figure 6.

To demonstrate the model’s operation with the city block learning rule, the dissimilarity data reported by Gati and Tversky (1982, Table 1, upper-triangular half) was employed. This data consists of the mean dissimilarity ratings of 29 subjects on a 20 point scale, for all 120 possible pairings of a 16 element stimulus set of drawings depicting flowers in flower pots. As shown in Figure 5.4, the drawings differ solely with respect to the depiction of the ‘elongation’ of the flowers and the ‘form’ of the pots. Both of these features have four possible realisations, labelled ‘a’, ‘b’, ‘c’ and ‘d’ for the form of the pots, and ‘p’, ‘q’, ‘r’ and ‘s’ for the elongation of the flowers, which are

independently and exhaustively varied across the stimulus set. Because of this method of generation, it seems reasonable to assume that the stimulus dimensions are appropriately regarded as psychologically separable.

The fact, however, that the feedback provided to the model is assumed to consist of similarity measures necessitates a transformation of the provided dissimilarity measures. Following Shepard (1962b, p. 232), this was accomplished simply by reflecting each dissimilarity value about the mid-point of the measurement scale. These similarity values were then normalised to lie between zero and one. The resultant similarity matrix is given in Table 5.2.

Table 5.2. Rescaled flower pot similarity matrix

	ap	aq	ar	as	bp	bq	br	bs	cp	cq	cr	cs	dp	dq	dr	ds
ap	1.000	0.630	0.555	0.535	0.735	0.290	0.305	0.170	0.620	0.290	0.265	0.145	0.605	0.200	0.190	0.135
aq	0.630	1.000	0.740	0.650	0.295	0.650	0.475	0.295	0.215	0.640	0.395	0.255	0.190	0.545	0.305	0.210
ar	0.555	0.740	1.000	0.750	0.260	0.460	0.670	0.475	0.240	0.405	0.635	0.420	0.185	0.290	0.555	0.395
as	0.535	0.650	0.750	1.000	0.185	0.330	0.390	0.700	0.155	0.260	0.310	0.660	0.130	0.245	0.270	0.570
bp	0.735	0.295	0.260	0.185	1.000	0.635	0.550	0.495	0.795	0.395	0.420	0.310	0.700	0.295	0.305	0.200
bq	0.290	0.650	0.460	0.330	0.635	1.000	0.755	0.680	0.385	0.785	0.585	0.460	0.290	0.690	0.425	0.300
br	0.305	0.475	0.670	0.390	0.550	0.755	1.000	0.775	0.365	0.555	0.785	0.575	0.230	0.445	0.670	0.445
bs	0.170	0.295	0.475	0.700	0.495	0.680	0.775	1.000	0.360	0.475	0.535	0.775	0.200	0.275	0.420	0.720
cp	0.620	0.215	0.240	0.155	0.795	0.385	0.365	0.360	1.000	0.610	0.565	0.515	0.800	0.430	0.400	0.335
cq	0.290	0.640	0.405	0.260	0.395	0.785	0.555	0.475	0.610	1.000	0.795	0.610	0.450	0.750	0.590	0.450
cr	0.265	0.395	0.635	0.310	0.420	0.585	0.785	0.535	0.565	0.795	1.000	0.715	0.420	0.570	0.735	0.605
cs	0.145	0.255	0.420	0.660	0.310	0.460	0.575	0.775	0.515	0.610	0.715	1.000	0.305	0.450	0.565	0.795
dp	0.605	0.190	0.185	0.130	0.700	0.290	0.230	0.200	0.800	0.450	0.420	0.305	1.000	0.625	0.580	0.570
dq	0.200	0.545	0.290	0.245	0.295	0.690	0.445	0.275	0.430	0.750	0.570	0.450	0.625	1.000	0.795	0.710
dr	0.190	0.305	0.555	0.270	0.305	0.425	0.670	0.420	0.400	0.590	0.735	0.565	0.580	0.795	1.000	0.775
ds	0.135	0.210	0.395	0.570	0.200	0.300	0.445	0.720	0.335	0.450	0.605	0.795	0.570	0.710	0.775	1.000

The flower pot model consisted of 16 units in each of the input, exemplar and feedback layers, in accordance with the number of stimuli, but was identical to the colour model with regard to the number of units in the internal representation layer, the initial placement of stimuli in the representational space, the choice of parameter values, and the number of trials employed.

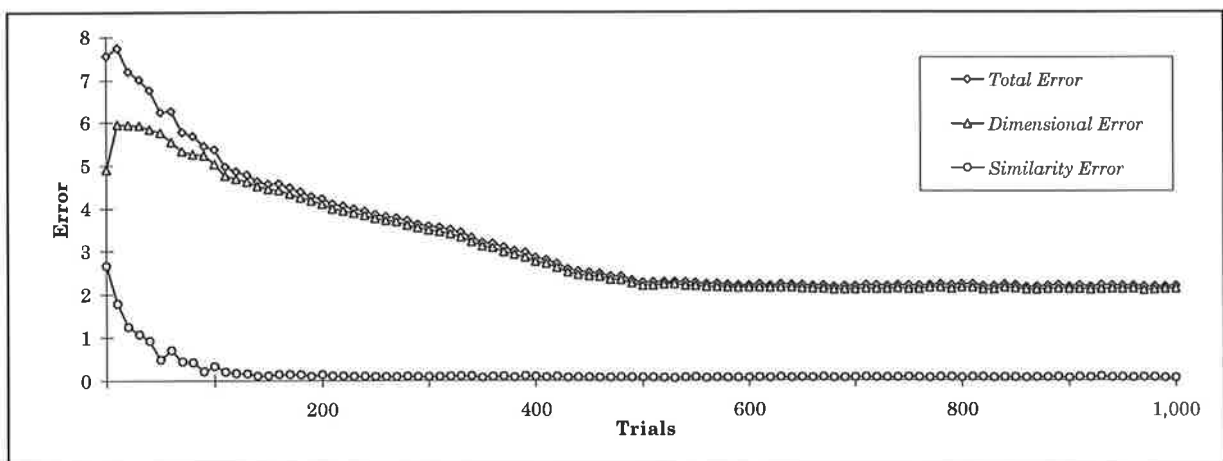


Figure 5.5. The pattern of change of the three error measures over 1,000 trials for the flower pot model.

The pattern of change of the three error measures across the 1,000 trials is shown in Figure 5.5. Once again, the similarity error measure exhibits a rapid decrease towards a negligible value which is maintained, indicating that the required patterns of inter-stimulus similarity have been

accommodated by the learned internal representational structure.

The decomposition of the dimensional error shown in Figure 5.6 indicates that the final learned internal representation consists only of stimulus dimensions 1 and 2. The resultant two-dimensional psychological space is shown in Figure 5.7. With reference to the stimulus set depicted in Figure 5.4, the appropriateness of the learned representation is evident. The stimulus dimensions developed by the model correspond precisely to the stimulus dimensions employed to construct the stimulus set: those of flower elongation and pot form. Furthermore, the relative spacing of the learned representations on this 4x4 grid seem entirely reasonable. For example, in relation to the form of the flower pot, the relatively larger distance between types 'a' and 'b' could be seen to correspond to the relatively greater transition of form involved in the departure from squareness. Similarly, the difference between flower elongations 'q' and 'r' seems smaller than the difference between 'r' and 's', and this is also evident in the learned psychological space representations.

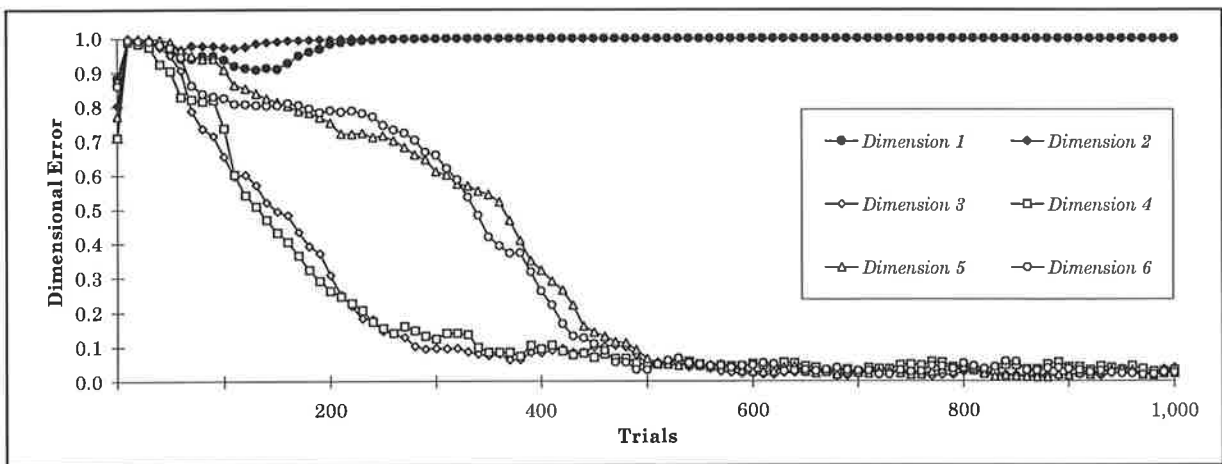


Figure 5.6. The breakdown of dimensional error across the 6 component stimulus dimensions for the flower pot model.

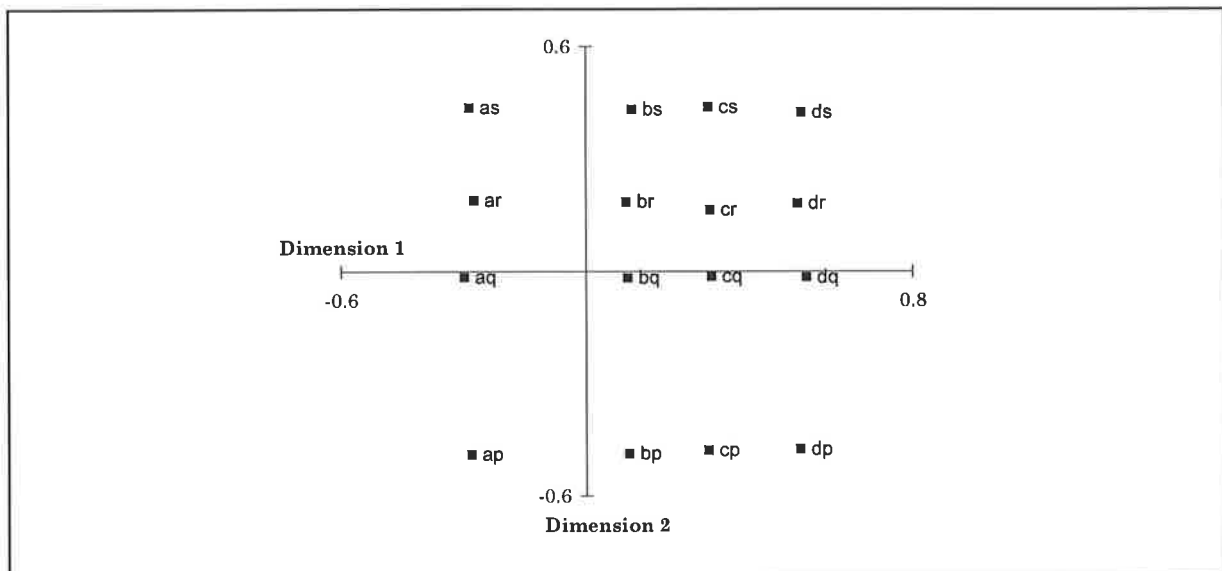


Figure 5.7. The final two-dimensional internal representation developed by the flower pot model.

---

## 5.2. Evaluation Of The Model

An evaluation of the model must consider at least three aspects of its apparent ability to learn appropriate psychological space internal representations. First, it is necessary to examine the effects various settings of the two free parameters - the learning rate parameter  $\lambda_c$ , and the dimensionality reduction parameter  $\beta$  - have upon the learned representations. Secondly, because the model can assume that the multidimensional stimuli it encounters are psychologically separable by employing the City-block distance metric in psychological space, the serious doubts raised by Arabie (1991, see also Hubert, Arabie & Hesson-Mcinnis 1992, Shepard 1974) concerning problems inherent in developing City-block multidimensional scaling algorithms within a gradient descent optimisation framework need to be addressed. Finally, it is important to examine the model's capabilities in relation to entrapment in local minima, a more widely appreciated limitation of gradient based optimisation strategies.

### 5.2.1. Sensitivity To Parameter Values

From the outset, it should be noted that the setting of the learning rate parameter  $\lambda_c$  is far less problematic than the dimensionality reduction parameter  $\beta$ . The component of the learning rule derived through gradient descent on the similarity error measure is essentially an iterative correction procedure. Such procedures are generally accepted to be relatively insensitive to the precise value of the learning rate parameter, providing this value is not so large as to result in oscillatory or degenerative behaviour. Values such as 0.01, 0.05, 0.1 are commonly employed without recourse to further principled justification on the grounds of this insensitivity. Simulations of the model confirm that the learned internal representations are extremely insensitive to the precise value of the learning rate parameter, although the rate of convergence can be significantly increased by the adoption of larger values.

The setting of the dimensionality reduction parameter, however, would appear potentially to have a greater influence on the satisfactory operation of the model. Indeed, extreme dimensionality reduction values will clearly prevent the model from deriving appropriate internal representations. An excessively small dimensionality reduction parameter value will hinder the model's ability to discern reliably the relative representational contribution made by each stimulus dimension, whilst an overly large value will prevent the dimensional error measure from identifying stimulus dimensions which fail to make significant representational contributions. In terms of the sensitivity of the model's performance to the dimensionality reduction parameter, therefore, it is important to develop some appreciation of the actual values of these upper and lower bounds which correspond to 'overly large' and 'excessively small' values.

To this end, the flower pot simulation described above was repeated with the learning rate parameter maintaining the value 0.1, but with the dimensionality reduction parameter assuming

the values 5 and 15. The significant variation of the dimensional error measure affected by this spread of values is evident by recalling the role of  $\beta$  in parameterising a negative exponential decay function. As shown in Figure 5.8., with a dimensionality reduction parameter value of 5, the total error stabilised at a minimum value after about 200 trials, with both the similarity and dimensional errors decreasing rapidly during this period.

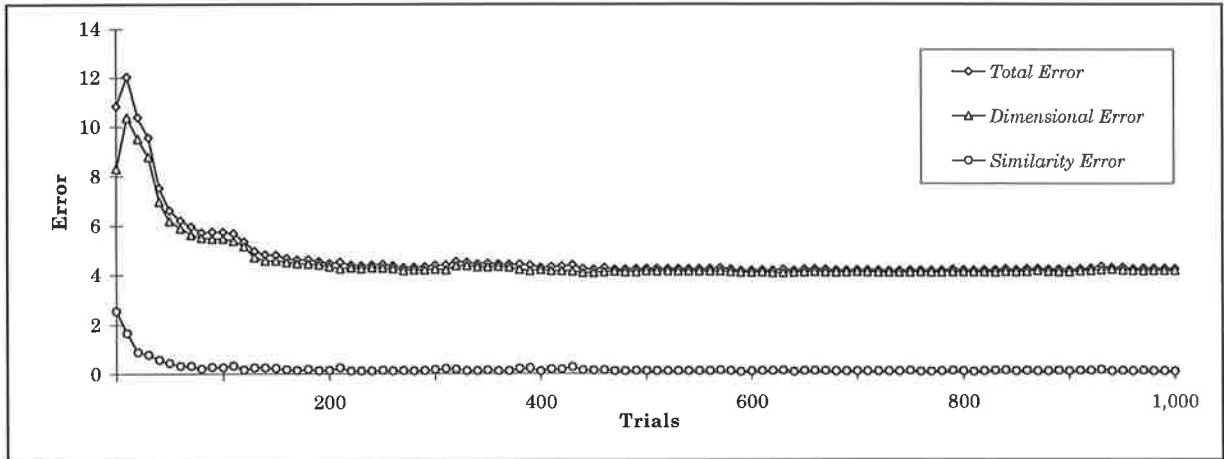


Figure 5.8. The pattern of change of the three error measures over 1,000 trials for the flower pot model with a dimensionality reduction parameter value of 5.

The representational space formed by the model in reaching this stable state was once again two-dimensional, as is evident from the patterns of dimensional error across the component stimulus dimensions shown in Figure 5.9.

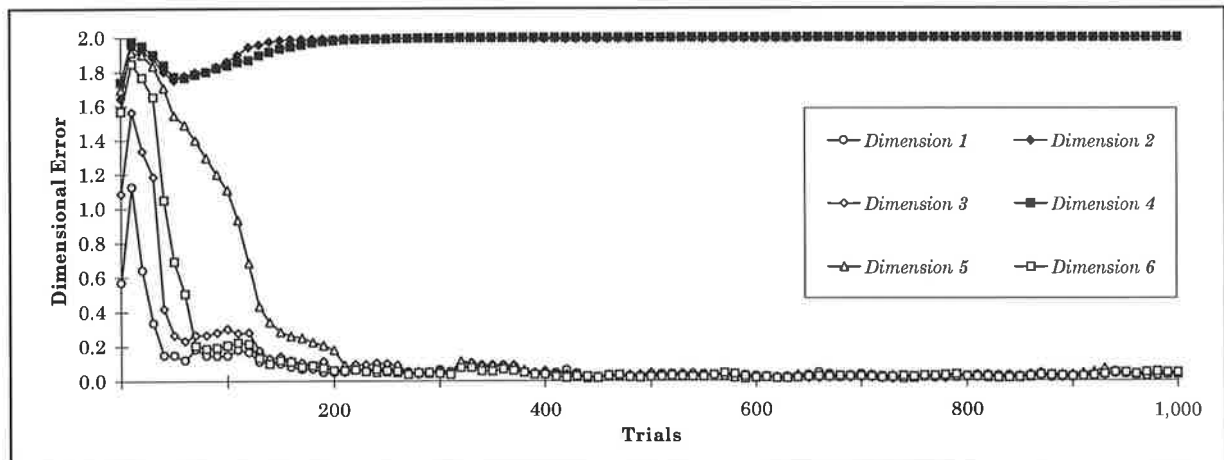


Figure 5.9. The breakdown of dimensional error into the 6 component stimulus dimensions for the flower pot model with a dimensionality reduction parameter value of 5.

The two-dimensional representational structure developed by the model is shown in Figure 5.10. Appreciation of the appropriateness of this result hinges upon a comprehension of the fact that, from the model's perspective, representational structures are unique only up to the arbitrary decisions made regarding their placement and conferred directionality, as required for graphical depiction. Thus, any derived representational structure can legitimately be manipulated by interchanging axes and by reversing the order in which the represented stimuli are displayed. Given the isomorphism between representational structures attainable through the application of

any sequence of such transformations, it is clear that Figure 5.7 and Figure 5.10 are merely different graphical manifestations of the same psychological space representation.

As shown in Figure 5.11, with a dimensionality reduction parameter value of 15, despite the fact that the similarity error is essentially zero from trial 200 onwards, the total error and dimensional error measures do not stabilise until approximately 600 trials have elapsed.

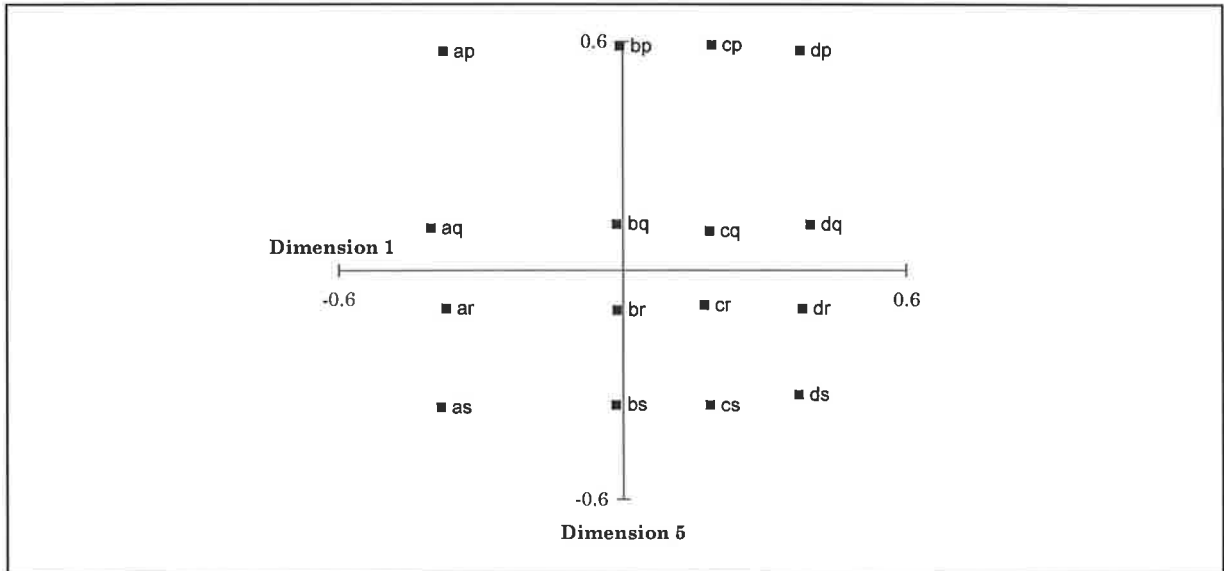


Figure 5.10. The final two-dimensional internal representation developed by the flower pot model with a dimensionality reduction parameter value of 5.

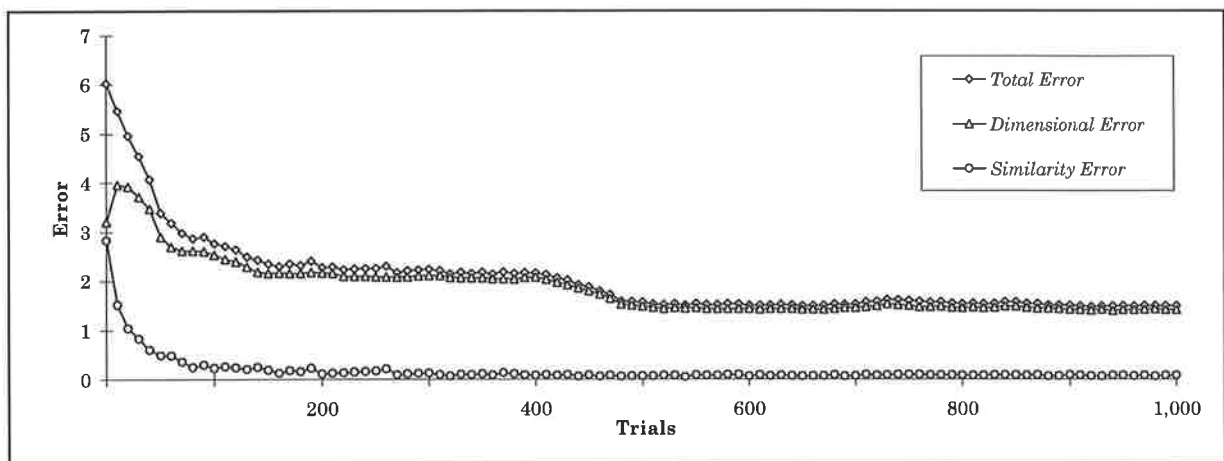


Figure 5.11. The pattern of change of the three error measures over 1,000 trials for the flower pot model with a dimensionality reduction parameter value of 15.

This delay is caused by the removal of one the stimulus dimensions from the representational structure between trials 300 and 500, as shown in the breakdown of the dimensional error across the component stimulus dimensions in Figure 5.12.

Once this change has been affected, however, the psychological space developed by the model is again two-dimensional, and contains the pattern of representations shown in Figure 5.13. Clearly, this derived set of internal representations does not significantly differ from those derived with dimensionality reduction parameter values of 5 and 10, and constitutes an appropriate psychological space representation of the stimulus set.

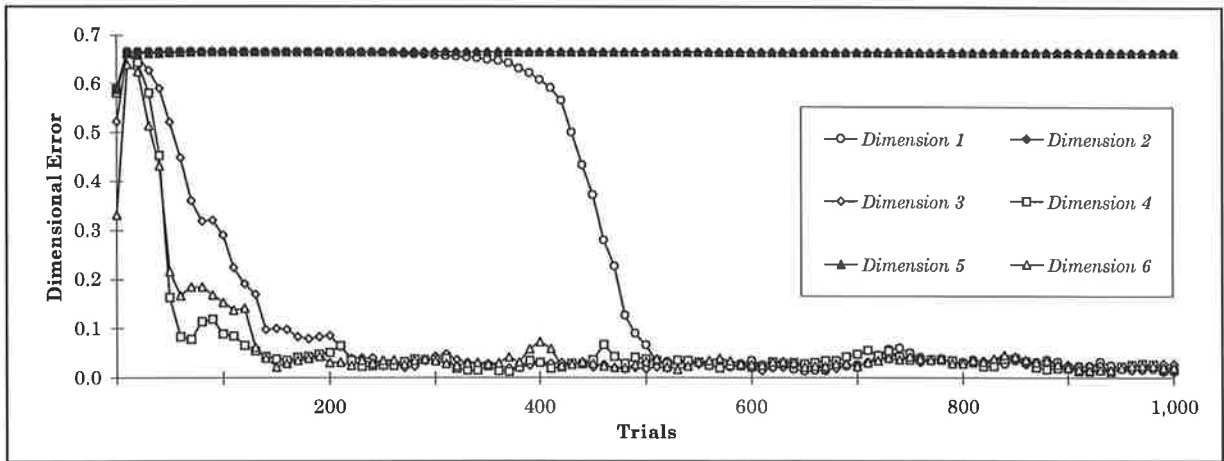


Figure 5.12. The breakdown of dimensional error into the 6 component stimulus dimensions for the flower pot model with a dimensionality reduction parameter value of 15.

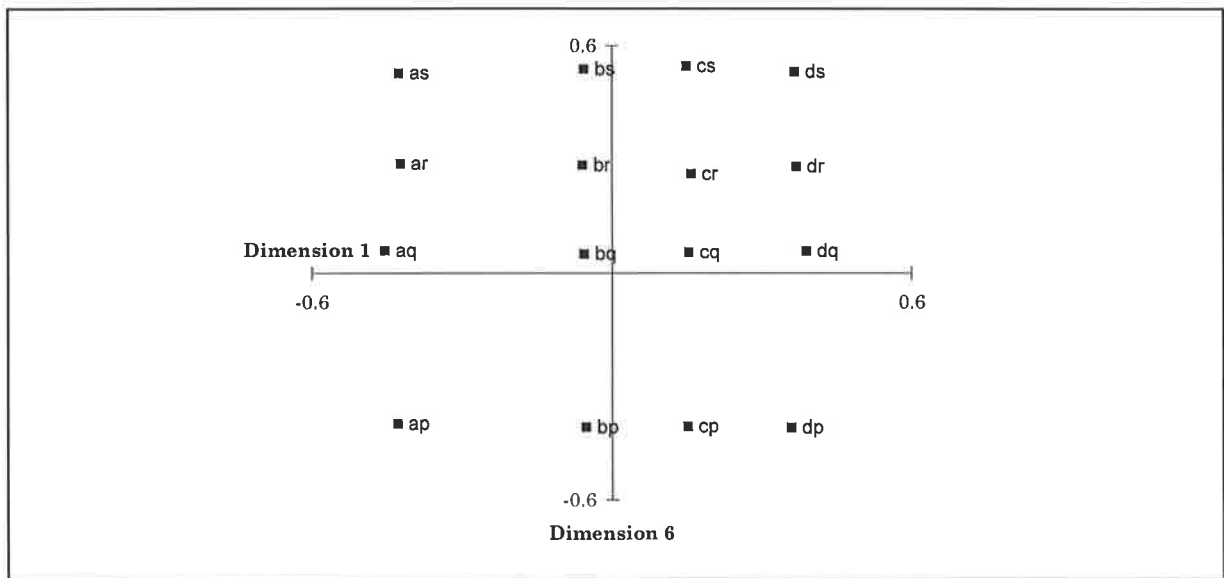


Figure 5.13. The final two-dimensional internal representation developed by the flower pot model with a dimensionality reduction parameter value of 15.

Given the ability of the model to develop the same representation across the parameter regime described, it seems reasonable to suggest that the appropriateness of this representation is not particularly dependent upon the value of the dimensionality reduction parameter. Of course, repeated simulation does not amount to a general demonstration of the relationship between learned representations and parameter settings. Nevertheless these demonstrations do show that, at least in relation to the flower pot similarity data, the representations being learned can be regarded as being a robust outcome of the application of multidimensional scaling principles, rather than parameter specific idiosyncrasies of the model. Since each of these additional simulations of the flower pot model involved different random initial representational configurations, a second inference which could be drawn is that the potential problem of sensitivity to initial conditions in connectionist models (see Kolen & Pollack 1991) seems not to be particularly prevalent within the model. The validity with which this conclusion can be generalised to other data sets probably depends, at least in part, on the nature of those data sets - and particularly upon the underlying

presence of an error ‘elbow’, as discussed in Section 3.1.5. Therefore, examining the ability of the model to develop the same representational structure from other data sets, across a wide range of dimensionality reduction parameter values, would appear to be a worthwhile exercise.

In any case, extending the model to allow the adaptive modification of the dimensionality reduction parameter during the model’s operation offers a means of further reducing the reliance of representational outcomes on parameter values. By initially setting the dimensionality reduction parameter to a small value, the extent to which it should be increased is readily gauged from the pattern of change of the dimensional error measure. Specifically, the dimensionality reduction parameter value is appropriately increased by a small amount whenever the dimensional error stabilises, until a situation is reached in which further increases in the dimensionality reduction parameter cease to induce a further decrease in the dimensional error. This type of adaptive approach to parameter setting is widely adopted in general connectionist modelling (eg. Weigend, Rumelhart & Huberman 1991, see Haykin 1994 for an overview), has previously been pursued with regard to multidimensional scaling algorithms (eg. Kruskal 1964b), and would certainly constitute a worthwhile extension of the model.

### 5.2.2. Overcoming Separable Stimulus Difficulties

Shepard’s (1974) discussion of the problems and prospects of multidimensional scaling algorithms places considerable emphasis on difficulties evident in multidimensional scaling using the City-block metric. More recently, these difficulties have been emphatically restated and significantly developed by Arabie (1991) and Hubert, Arabie and Hesson-Mcinnis (1992). Clearly, given that the model has the scope to represent psychologically separable stimuli through employing the City-block metric in psychological space, it is important that these concerns are addressed.

In essence, the claimed difficulties inherent in multidimensional scaling over the City-block metric are founded upon apparent inadequacies in the gradient-descent optimisation framework. Specifically, Hubert et. al. (1992) argue that:

“gradient strategies in a city-block application lead to the satisfaction of necessary conditions that are simply too weak to be used as the core of an adequate optimisation algorithm” (p. 212).

The theoretical arguments presented in support of this claim, however, assume that the multidimensional scaling algorithm operates in a representational space of fixed dimensionality. Under this scenario it is demonstrated (Hubert et. al. 1992, pp. 216-217) that the gradient descent optimisation approach does not sufficiently constrain the set of appropriate representational configurations, in the sense that a large number of configurations which do not constitute appropriate ‘solutions’ are stable states of the optimising system.



It is, however, at least plausible to argue that the model described here does not suffer from such shortcomings. The requirement of minimal dimensionality with regard to the derived representation, as formalised by the dimensional error measure, constitutes precisely the type of additional constraint which the analysis of Hubert et. al. (1992) suggests is needed. The evaluation of this possibility is most directly achieved by applying the model to the type of task presented as a concrete demonstration of the theoretically suggested deficiencies by Hubert et. al. (1992).

This task involves examining the ability of a multidimensional scaling algorithm to recover a pre-determined representational structure which, assuming the City-block distance metric in the representational space, has been employed to generate similarity data. Hubert et. al. (1992) provide convincing demonstrations of the inability of traditional, fixed dimension, gradient descent based multidimensional scaling algorithms to reveal the representational structure which is, by its method of generation, known to underlie such similarity data.

The particular choice of representational structure used by Hubert et. al. (1992) is an evenly spaced 5x5 square lattice with stimuli randomly placed at 15 of the 25 available positions. The actual configuration examined in detail by Hubert et. al. (1992), with normalised axes, is shown in Figure 5.14. The reason for the scaling of the configuration will become clear in Chapter 7, but, for the moment, it suffices to note that this alteration does not in any way modify the reconstruction problem. There is no reason to believe that the choice of scale adopted by Hubert et. al. (1992) is anything other than arbitrary, and the normalisation evident in Figure 5.14 does not alter the representational configuration in any other way.

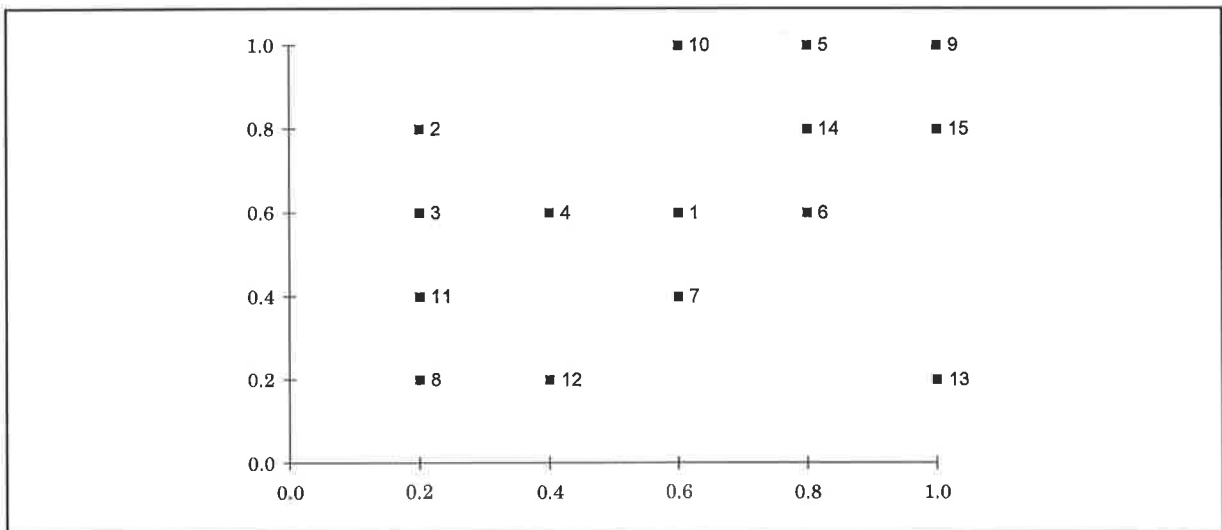


Figure 5.14. The normalised initial configuration of 15 points employed by Hubert et. al. (1992). Adapted from Hubert et. al. (1992), Figure 1a.

The similarity matrix provided to the model, given in Table 5.3, was derived from the pre-determined representational configuration. This was done by determining the City-block distance between each pair of stimuli, and then applying this distance as the argument to the exponential decay function given in Equation 4.4, as assumed by the model in accordance with the Universal

Table 5.3. The lattice reconstruction similarity matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.000	0.549	0.670	0.819	0.549	0.819	0.819	0.449	0.449	0.670	0.549	0.549	0.449	0.670	0.549
2	0.549	1.000	0.819	0.670	0.449	0.449	0.449	0.549	0.368	0.549	0.670	0.449	0.247	0.549	0.449
3	0.670	0.819	1.000	0.819	0.368	0.549	0.549	0.670	0.301	0.449	0.819	0.549	0.301	0.449	0.368
4	0.819	0.670	0.819	1.000	0.449	0.670	0.670	0.549	0.368	0.549	0.670	0.670	0.368	0.549	0.449
5	0.549	0.449	0.368	0.449	1.000	0.670	0.449	0.247	0.819	0.819	0.301	0.301	0.368	0.819	0.670
6	0.819	0.449	0.549	0.670	0.670	1.000	0.670	0.368	0.549	0.549	0.449	0.449	0.549	0.819	0.670
7	0.819	0.449	0.549	0.670	0.449	0.670	1.000	0.549	0.368	0.549	0.670	0.670	0.549	0.549	0.449
8	0.449	0.549	0.670	0.549	0.247	0.368	0.549	1.000	0.202	0.301	0.819	0.819	0.449	0.301	0.247
9	0.449	0.368	0.301	0.368	0.819	0.549	0.368	0.202	1.000	0.670	0.247	0.247	0.449	0.670	0.819
10	0.670	0.549	0.449	0.549	0.819	0.549	0.549	0.301	0.670	1.000	0.368	0.368	0.301	0.670	0.549
11	0.549	0.670	0.819	0.670	0.301	0.449	0.670	0.819	0.247	0.368	1.000	0.670	0.368	0.368	0.301
12	0.549	0.449	0.549	0.670	0.301	0.449	0.670	0.819	0.247	0.368	0.670	1.000	0.549	0.368	0.301
13	0.449	0.247	0.301	0.368	0.368	0.549	0.549	0.449	0.449	0.301	0.368	0.549	1.000	0.449	0.549
14	0.670	0.549	0.449	0.549	0.819	0.819	0.549	0.301	0.670	0.670	0.368	0.368	0.449	1.000	0.819
15	0.549	0.449	0.368	0.449	0.670	0.670	0.449	0.247	0.819	0.549	0.301	0.301	0.549	0.819	1.000

The model consisted of 15 units in each of the input, exemplar and teacher layers, and again had 6 units in the internal representation layer. The learning rate parameter was set to 0.1, whilst the dimensionality reduction parameter took the value 10. The value of the three error measures across 1,000 trials is shown in Figure 5.15 below. Whilst the similarity error rapidly falls to a negligible value, the dimensional and total error measures, after reaching an early plateau, experience a further significant decrease around the 400 trial stage.

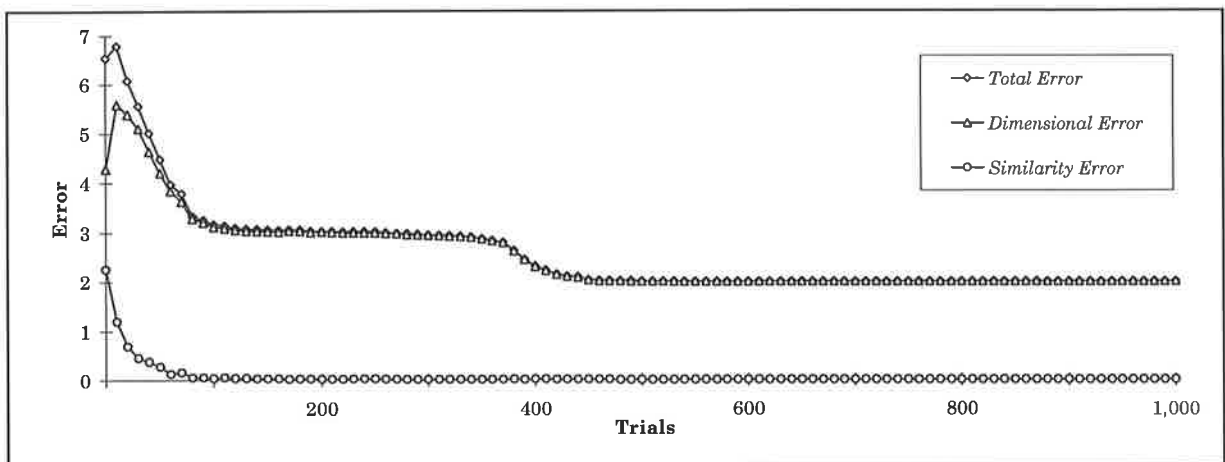


Figure 5.15. The pattern of change of the three error measures over 1,000 trials for the lattice reconstruction models.

The cause of this pattern of change is, once again, evident from an examination of the individual stimulus dimensions, as presented in Figure 5.16. The belated removal of a stimulus dimension from the representation being developed by the model results in the further decrease in both error measures, and results in a final representational structure which is clearly two-dimensional.

This two-dimensional representational structure is graphically depicted in Figure 5.17. With reference to the pre-determined initial configuration shown in Figure 5.14, and recalling the allowable manipulations of derived representations, it is evident that the model has successfully

recovered the initial configuration. In particular, if the direction conferred on the ordinate is reversed, and the two axes are subsequently interchanged, then the representational structure learned by the model coincides with the configuration from which the similarity matrix was derived.

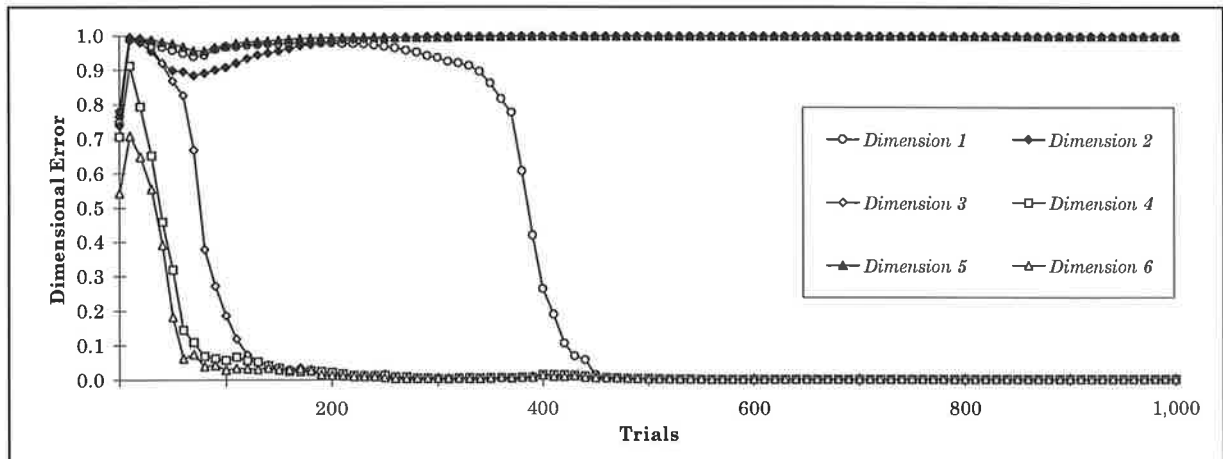


Figure 5.16. The breakdown of dimensional error into the 6 component stimulus dimensions for the lattice reconstruction models.

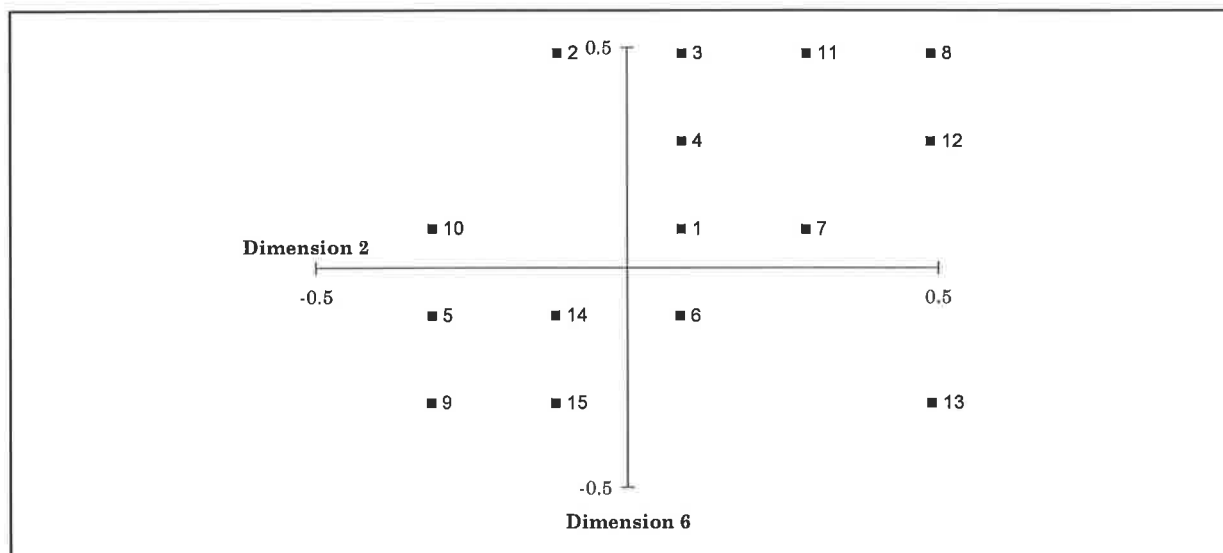


Figure 5.17. The final two-dimensional internal representation developed by the lattice reconstruction model.

The ability of the model to reconstruct the appropriate lattice configuration provides a basis on which to claim that the model developed in this chapter reliably and validly applies gradient descent optimisation principles to multidimensional scaling representational principles, despite the fact that it operates under the City-block distance metric. It is difficult to provide a detailed explanation of the model's success with regard the claimed metric specific difficulties, although clearly it is related to the form of the dimensional error component of the total error measure.

### 5.2.3. Entrapment In Local Minima

The adoption of a gradient-descent based optimisation strategy also introduces the possibility of inappropriate psychological spaces being derived because of the presence of local minima in the total error measure. In this context Intrator and Edelman (1996) suggest that the deterministic

annealing multidimensional scaling approach developed by Hofmann and Buhmann (1994, see also Klock & Buhmann 1997) may be superior to gradient-descent approaches. Simple explorations provide preliminary support for this view, indicating that the large numbers of inter-point distance constraints which constitute the similarity error measure may be conducive to the creation of local minima.

For example, consider the geometric locations of three stimulus-points,  $x$ ,  $y$  and  $o$ , in a one-dimensional space, required to meet the following similarity constraints:  $x$  must be 1 unit from  $o$ ,  $y$  must be  $1\frac{1}{2}$  units from  $o$ , and  $x$  and  $y$  must be  $\frac{1}{2}$  a unit from each other. Without loss of generality, the location of stimulus point  $o$  may be fixed at the origin. Clearly, the pattern of similarities is attainable in a one dimensional space, with  $(x, y)$  values of  $(1, 1\frac{1}{2})$  or  $(-1, -1\frac{1}{2})$  achieving an error measure of zero.

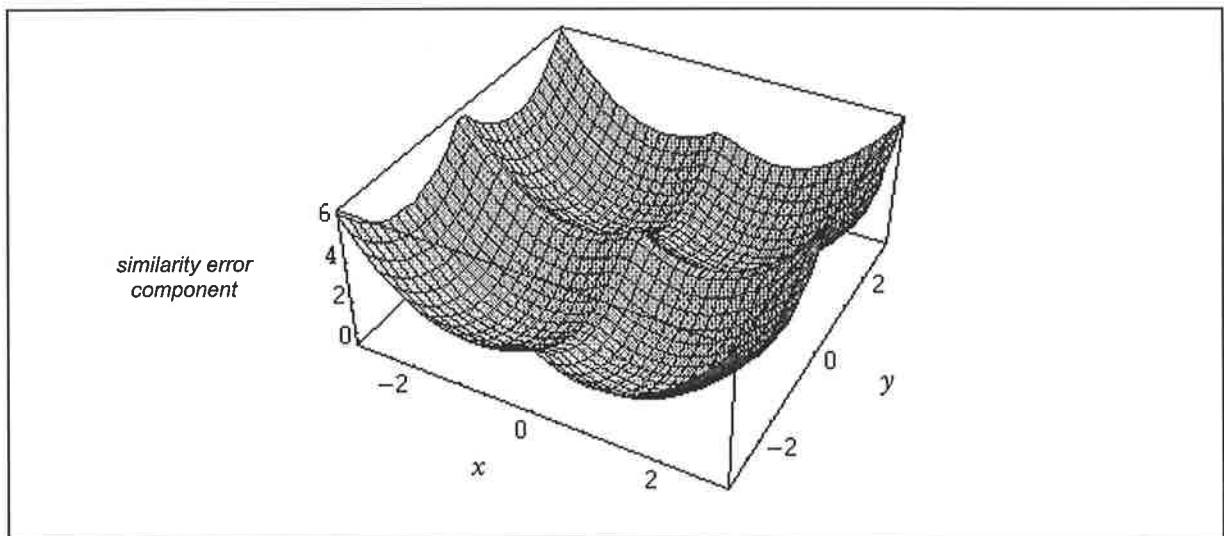


Figure 5.18. The error surface implied by the similarity of  $x$  and  $y$  in relation to  $o$ .

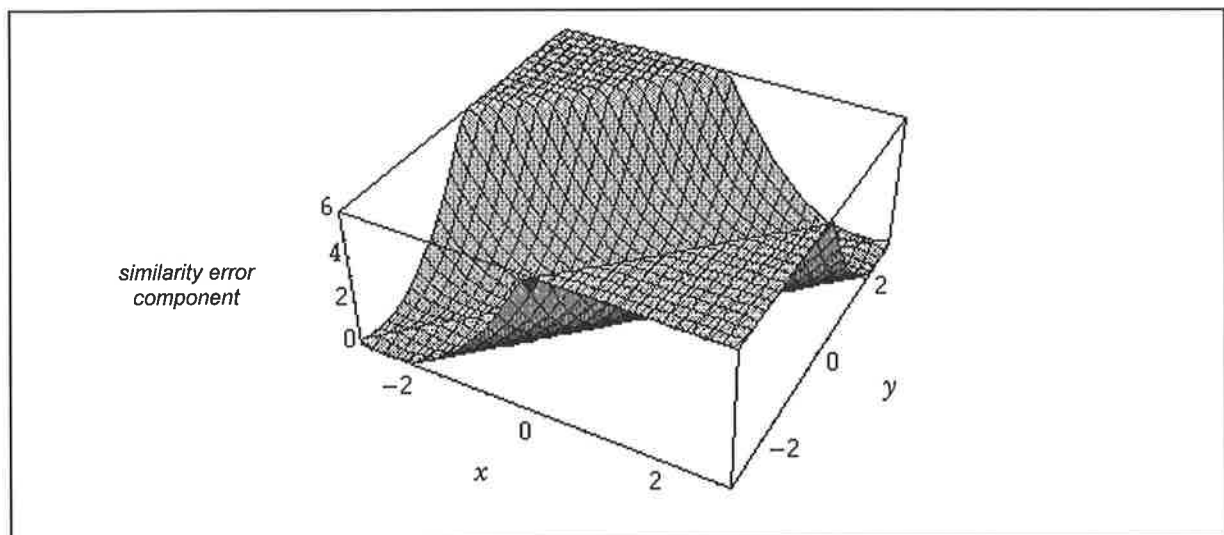


Figure 5.19. The error surface implied by the similarity of  $x$  and  $y$ , thresholded at an error value of 6 to assist graphical depiction.

To understand the nature of the error measure, as a function of the location of  $x$  and  $y$ , it is useful to decompose the measure into two parts. First, consider the constraints involving the

distance from  $o$  to both  $x$  and  $y$ , as shown on the left of Figure 5.18. This error surface has four points at which the error is zero, namely  $(1, \frac{1}{2})$ ,  $(1, -\frac{1}{2})$ ,  $(-1, \frac{1}{2})$ ,  $(-1, -\frac{1}{2})$ , corresponding to the four ways in which  $x$  may be 1 unit from the origin whilst  $y$  is  $\frac{1}{2}$  units from the origin. Secondly, consider the error surface implied by the requirement that  $x$  and  $y$  be separated by  $\frac{1}{2}$  a unit, as shown in Figure 5.19. This surface has two parallel 'troughs' on the lines  $y = x - \frac{1}{2}$  and  $y = x + \frac{1}{2}$ .

The similarity error surface for this three point problem is the sum of these component error surfaces, and is shown in Figure 5.20. The important point to note regarding this surface is the presence of local minima which, within a gradient-descent framework, are the basins of attraction for a significant proportion of possible initial values of  $x$  and  $y$ .

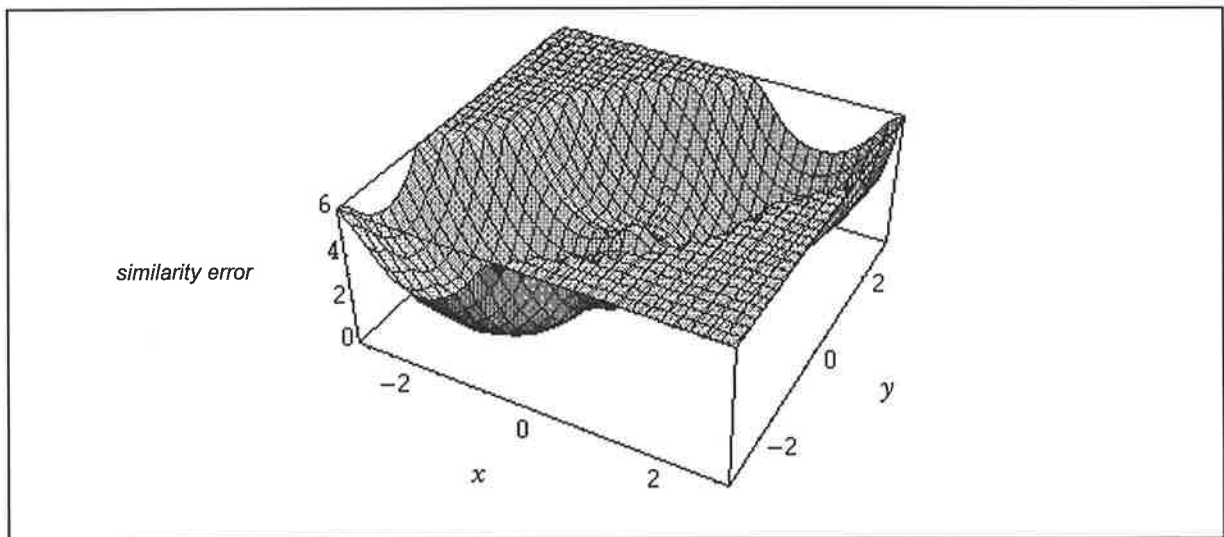


Figure 5.20. The combined error surface for the three point problem. Once again, the surface has been thresholded at an error value of 6 to assist graphical depiction.

This demonstration suggests that the form of similarity error which characterises gradient-descent based multidimensional scaling is certainly not immune to the presence of local minima. Thus, the potential problem of finding sub-optimal solutions using gradient-descent optimisation techniques through entrapment in these local minima must be addressed. Indeed, such difficulties are widely asserted (eg. Arabie 1991, Shepard 1974) to be particularly prevalent in the one-dimensional case. As such, the possibility arises of the model encountering difficulties when deriving a psychological space for a set of stimuli which is appropriately characterised in terms of a single psychological dimension.

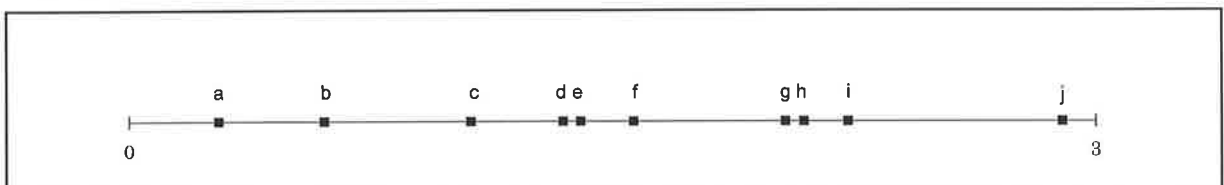


Figure 5.21. The initial random placement of the ten stimuli on the number line.

To examine this possibility, a one-dimensional reconstruction task was developed by randomly placing ten stimuli (labelled with the letters 'a' through 'j') on a number line extending from the value zero through the value three, as shown in Figure 5.21.

The model was given ten units in the input, exemplar, and feedback layers, and again employed six internal representation units. The dimensionality reduction parameter took the value 10, whilst the learning rate was set to 0.05 rather than 0.1, primarily to provide some demonstration of the model's asserted insensitivity to the precise value of this parameter. Figure 5.22 displays the behaviour of the three error measures over the first 500 trials. All three measures have stabilised by the conclusion of these trials, with the similarity error measure again showing a negligible value.

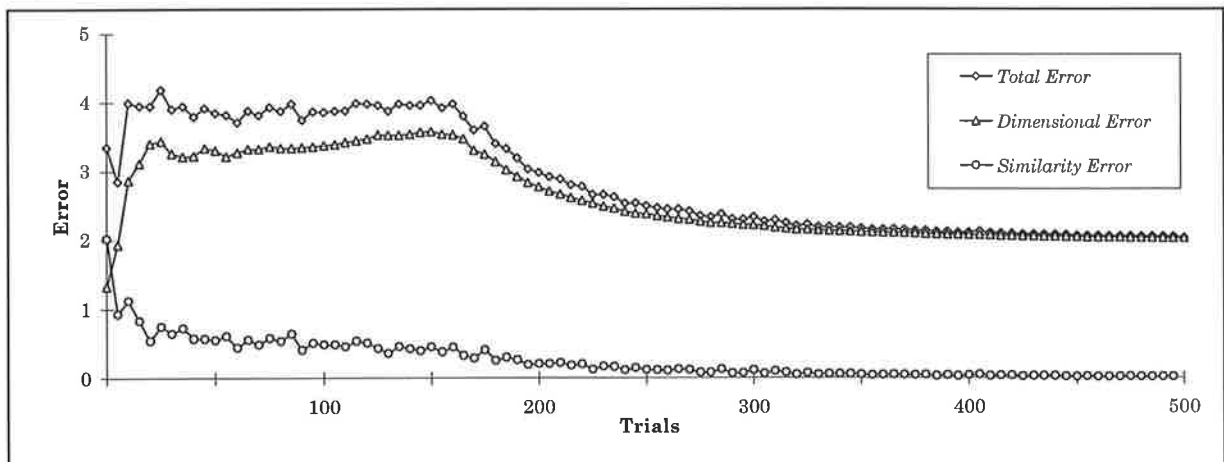


Figure 5.22. The pattern of change of the three error measures over 500 trials for the one-dimensional reconstruction model.

The breakdown of the dimensional error measure, as shown in Figure 5.23, indicates that the model's derived psychological space is one-dimensional, consisting solely of dimension 3.

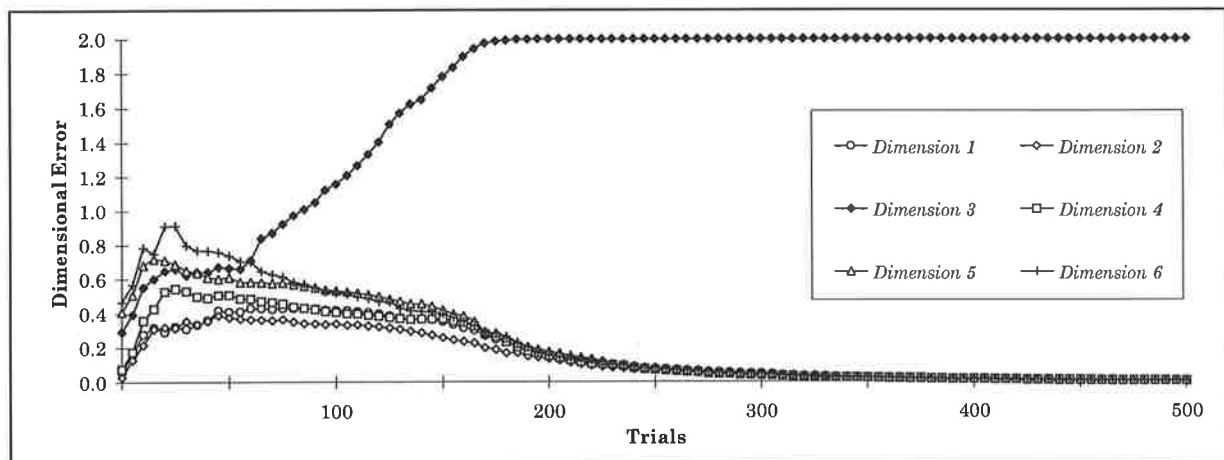


Figure 5.23. The breakdown of dimensional error into the 6 component stimulus dimensions for the one-dimensional reconstruction model.

Finally, the representational values assumed by the ten stimuli on dimension 3, as shown in Figure 5.24, accord, once their arbitrarily conferred direction is reversed, with the initial structure of Figure 5.21.

The fact that the model correctly reconstructs the one-dimensional locations of stimuli in this way strongly suggests that it has some considerable ability with regards the avoidance of local minima. In relation to this avoidance, Shepard (1974) has argued that the inclusion of mechanisms

for dimensionality determination within multidimensional scaling algorithms confer considerable advantages. Specifically, it is argued that, in representational spaces of a dimensionality which is higher than ultimately proves necessary, the ability of the model to avoid local minima is significantly enhanced. Intuitively, this argument suggests that a local minimum in a representational space of a certain dimensionality is unlikely to remain a local minimum in spaces of higher dimensionality, and, therefore, the application of gradient descent optimisation principles in these higher dimensional spaces may serve to appropriately position the various stimuli before the 'descent' into fewer dimensions is made.



Figure 5.24. The final one-dimensional internal representation developed by the one-dimensional reconstruction model.

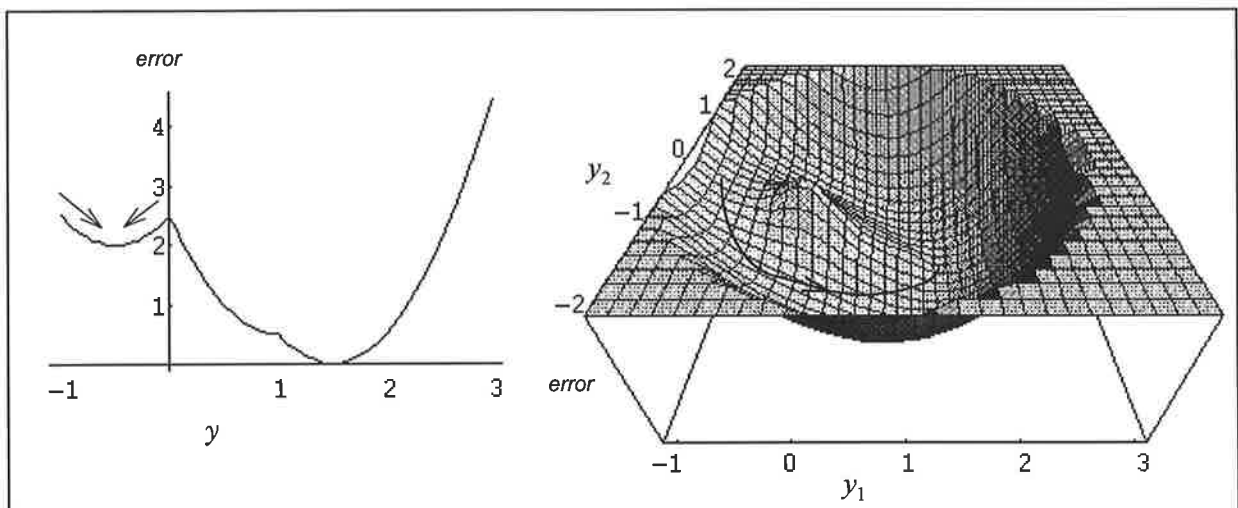


Figure 5.25. The one-dimensional and two-dimensional error surfaces for the three point problem in which both  $o$  and  $x$  are fixed.

Simple preliminary examinations of this intuition tend to provide encouraging affirmation and are graphically suggestive. Consider for example, the three point problem described above in which  $x$  must be 1 unit from  $o$ ,  $y$  must be  $1\frac{1}{2}$  units from  $o$ , and  $x$  and  $y$  must be  $\frac{1}{2}$  a unit from each other. To allow graphical depiction suppose, in this instance, that not only is  $o$  fixed at the origin, but that  $x$  is fixed at  $+1$ . The similarity error function for the point  $y$ , if restricted to one dimension, is shown on the left of Figure 5.25. As indicated by the arrows, the application of gradient-descent to any negative starting position for  $y$  results in the generation of an inappropriate solution through entrapment in a local minimum. If, however, a solution is initially sought in two-dimensions through locating  $o$  at  $(0,0)$  and  $x$  at  $(1,0)$ , then an error surface for the point  $y = (y_1, y_2)$  of the type shown on the right of Figure 5.25 is constructed. Within this surface, the local minima have become 'saddle-points' which are readily avoided by gradient-descent

optimisation principles. Indeed, the final location of  $\gamma$  for the two-dimensional case is the point  $(1\frac{1}{2}, 0)$  which corresponds to zero similarity error. The line on the right of Figure 5.25 indicates progress to this minimum for a case in which the initial  $\gamma_i$  is negative.

Once  $\gamma$  has reached this location, or one in the vicinity, the operation of the dimensional error would be expected to remove the surplus second dimension, since all three points assume the same value (ie. zero) on this dimension. In this way, the possibility of error minimisation in a higher-dimensional space seems to facilitate the derivation of an appropriate lower-dimensional representational structure. In this regard, the model developed in this chapter fares particularly well, since the units in the internal representation layer are not permanently removed once they coincide with the line of non-contribution, and remain available to temporarily act in representing structural information to allow the avoidance of a local minimum. For these reasons, it is suggested that the model developed in this chapter may be particularly impervious to entrapment in local minima.

The results presented in Sections 5.2.2 and 5.2.3 suggest, therefore, that the additional representational requirements provided by the dimensional error measure are sufficiently constraining to prevent the derivation of degenerate solutions, whether associated with metric-specific or local minima difficulties. A worthwhile topic of further study would be the development of a rigorous understanding of those features of the dimensional error which constitute these constraints. Such an understanding would have the immediate practical application of further refining the dimensional error measure. Although the measure of representational contribution and the broad features of the dimensional error measure have been developed in a principled manner, Equation 4.10 clearly represents only one of many functions exhibiting the necessary properties. A more detailed understanding of the interaction of the dimensional and similarity error measures might suggest alternative functions which would result in more appropriate measures of dimensional error. This issue is further discussed in Section 9.1.5.



## Chapter 6: The Internal Derivation Of Psychological Similarity

Whilst the model described in Chapter 4 appears to be capable of successfully learning psychological space internal representations, it does not constitute a reasonable model of the way in which humans acquire these representations. As was argued in Section 3.2.6, the reliance of the model on a pre-determined similarity matrix is inappropriate. Such indices of inter-stimulus similarity simply are not continually presented to humans by the external world. Therefore, if human mental representations are to be appropriately modelled by psychological spaces, a method must be found by which connectionist models are able to derive autonomously the measures of psychological similarity required for the construction of these spaces. Through considering the relationship between the human mind and the external world, this chapter arrives at a first tentative formalisation of such a method.

---

### 6.1. The Relationship Between The Mind And The World

The basic premise developed in this chapter is one with both a long-standing and an accepted currency within psychology: that the *structure* of mental representation reflects its *function*. More specifically, it is argued that, since one of the most important roles played by mental representational structures is to endow its possessor with some form of adaptive advantage, it follows that these structures must be significantly influenced by a set of representational constraints implicit in the external world. As summarised by Anderson (1990):

“the argument is that ... memory phenomena and the mechanisms that produce them are caused by the goals of the [human cognitive] system interacting with the structure of the environment” (p. 43)

Shepard’s (1990b) overview is that:

“We may look into that window on the mind as through a glass darkly, but what we are beginning to discern there looks very much like a reflection of the world” (p. 213)

It seems uncontroversial to assert that human and animal behaviour is influenced by the external world. Vickers’ (1979, pp. 88-89) description of an elementary organism with the capacity to move towards light on the basis of rudimentary processing of the sensory information it receives provides a simple demonstration of this relationship. Clearly, the behaviour of this organism is completely determined by presence or absence of light in the immediate environment.

Similarly, Simon (1981) provides a description of the path traversed by an ant moving along a beach in terms of the structure of the beach itself. In particular, the ant’s proclivity to avoid immediate obstacles is viewed as the adaptive ability which shapes the movement of the ant. Under this conception, the complexities evident in the path traversed by the ant, when viewed as a whole, are caused by the complexities of the beach environment in which the ant is situated. Simon (1981)

neatly summarises this relationship as follows:

“An ant, viewed as a behaving system, is quite simple. The apparent complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself” (p. 64)

Presumably, the means by which both Vickers' (1979) organism and Simon's (1981) ant produce behavioural action following environmental stimulation is through some form of innate neural information processing system. It also seems reasonable to assert that the evolution of such systems can profitably be viewed in terms of their provision of adaptive advantages. Guidance towards life-sustaining light, in the case of the organism, and the avoidance of immediate obstacles, in the case of the ant, are clearly both desirable behavioural capabilities. Given this scenario, the crudest application of evolutionary theory suggests that neural structures accommodating the appropriate behavioural abilities will become 'hard-wired' in future generations of the organisms and ants.

Such an adaptively driven process of internalisation, however, is equally well applied to information processing structures which do not simply serve to associate a stimulus with a response. There is also much to be gained from the development of neural structures which provide information about enduring, and possibly more abstract, regularities of the external world. As an example, consider Shepard's (1984, p. 422) discussion of the internalisation of the temporal night/day cycle in hamsters, as evidenced by their continuation of regular periods of activity and sleep within the artificial lighting conditions of a laboratory. The adaptive advantage of internalisations of this type arises from the endowment of an ability to produce appropriate behaviour in the absence of direct environmental stimulation. Indeed, in general, neural information processing structures which alleviate direct dependence upon sensory stimulation are desirable, if not indispensable, in a dynamic, unreliable, and often capricious world.

A recognition of the importance to human visual perception of enduring properties in the external environment underpins the notion of 'ecological optics' (see Gibson 1966, 1979). Within this theory, visual perception is conceived of as the process of extracting abstract invariants from the sensory stimulation provided by an external optic array. These invariants, in turn, are considered in terms of their various significances, or 'affordances', in relation to humans. A good example of evidence in favour of this conceptualisation of visual perception is found in empirical research involving the human perception of gaits (see, for example, Cutting 1981, Cutting, Proffitt & Kozlowski 1978). In essence, it is suggested that sensitivity to abstract invariants relating to joint movements underlies impressive human perceptual abilities in identifying other humans through the comprehending detection of their movement. Such abilities clearly afford adaptive advantages, and the detailed articulation of the responsible perceptual invariants in the form of joint movements - which are environmentally constrained through physical laws - reinforces the

potential of the ecological optics approach.

In an argument consonant with that of ecological optics, Shepard (1975, 1981b, 1984, 1987b, 1989, 1992a, 1994) proposes that evolution results in the development of neural systems which process information in a manner entirely constrained or determined by physical properties of the world. For example, human propensities to conceive the world as three-dimensional, locally Euclidean, subject to three degrees of both translational and rotational freedom, and having a unique downward direction conferred by gravity may all be cast as ones which are beneficial, if not necessary, for the successful comprehension of the external world. As Carlton and Shepard (1990a) summarise:

“If the internal process is to yield an adaptive outcome, its constraints cannot be determined by purely arbitrary limitations on the information processing system. Rather, the system must have been shaped, by natural selection and individual learning, to embody the most relevant constraints that govern the external world” (p. 128)

As with ecological optics, the structure of perceptual information processing systems is seen as being inextricably intertwined with the structure of the world. Shepard’s point of departure from the Gibsonian notion of ecological optics, however, arises when the neural systems responsible for such internalisations are subsequently viewed as being appropriately modelled as cognitive structures (see, in particular, Shepard 1984).

Whether or not the internalisation of the abstract structure of salient properties of external stimuli is profitably viewed as implying the existence of a cognitive system, or is considered simply a biological fact, it is clear that the internalisation process itself constitutes a fundamental departure from the pairing of stimuli and responses evident in Vickers’ (1979) organism and Simon’s (1981) ant. Rather than determining behaviour solely through simple associations, the environment is now viewed as a structured system, with abstract properties which are all-pervasive in their influence on human life.

A further fundamental difference between Simon’s (1981) ant and humans is developed by Vera and Simon (1993) through noting that the:

“ant does not need (and almost certainly does not have) a centralized and permanent representation of its environment” (p. 34)

and then suggesting that:

“higher organisms, however, appear to operate on more robust representations of the world than the ant ... This requires a significantly more complex representation than the ant’s, one that is more permanent and can be manipulated to abstract new information” (p. 35)

Essentially the same argument is advanced by Shepard (1984):

“higher organisms are not merely observers; they are active explorers and manipulators of their environment. If such exploration and manipulation is not

just random trial and error, it must be guided by some internal schema or hypothesis. At this point, a new type of function emerges that is related to perceptual and to motoric functions, but is not identical to either. I refer to ... the ability to remember, to anticipate, and to plan objects and events in their absence” (Shepard 1984, p. 421)

In other words, not only do properties of the world constrain perceptual processes, but properties of the various objects or stimuli within that world also affect human life. Successful interaction with the external environment requires both the ability to perceive that environment, and the ability to comprehend the properties and consequences of the myriad of objects encountered within that environment. Following the recognition of learning theorists (eg. Rescorla & Wagner 1972) “that it is the predictive or informational significance of events rather than mere contiguity that is the basis of learning” (Shepard 1992b, p. 420), it is clearly imperative that humans develop neural information processing structures which allow the prediction of the properties and consequences of external objects. Shepard (1988b) lucidly refers to the process of acquiring such abilities as the development of a ‘metric of functional equivalence’.

At this point, the evidence supporting the adoption of a cognitive conceptualisation of environmentally induced neural structures becomes overwhelming. Whatever stance is adopted in relation to the association of stimuli with responses, or the autonomous perception of invariant and fundamental properties of the world, the effects that the various properties of individual objects in the world have upon humans must be viewed in terms of the constraints they impose upon mental representational structures. The development of an understanding of the form of these constraints, therefore, constitutes a first step towards the specification of a mechanism which enables the modelling of the human acquisition of psychological spaces. Accordingly, the remainder of this section presents a survey of established psychological theory which attempts to formalise such an understanding.

### **6.1.1. The Rational Approach**

Anderson (1990, 1991, 1992) describes an attempt to formalise the relationship between the mind and the world through the application of the ‘rational man’ approach commonly identified with the field of economics. The essence of this approach involves considering human behaviour as rational, in the sense of realising the pursuit of human goals in a manner which, within a set of inherent information availability and processing capacity limitations, can be regarded as optimal. Consequently, under this view, the cognitive structures which underpin human behaviour are regarded as constituting optimal solutions to constrained representation problems. As Anderson (1992) argues:

“if we know that behavior is optimized to the structure of the environment and we also know what that optimal relationship is, then a constraint on mental mechanisms is that they must implement that optimal relationship” (p. 186)

Thus, the application of a rational approach to the development of a model of memory retrieval (Anderson 1990, 1992, Anderson & Schooler 1991) suggests that the probability of information recall from memory reflects the relative cognitive costs of retrieval, and the expected behavioural benefit resulting from the availability of the information. Similarly, a rational approach to modelling the categorisation process (Anderson 1990, 1991, 1992) dictates that conceptual structures constitute maximally informative groupings of a set of memory elements. Anderson (1990, 1992) also develops rational models of processes of causal inference and problem solving based on the same principle of adaptation through constrained optimisation.

The ability of cognitive models formulated under the rational approach to emulate empirical data is convincingly demonstrated in the extensive evaluation provided by Anderson and others (see, for example, Anderson 1990, 1991, 1992, Anderson & Schooler 1991, Oaksford & Chater 1994). The performance of such models is, necessarily, largely a function of the means by which human information processing constraints and the structure of the environment are formalised. Nevertheless, in many applications the specification of such formalisms, at least in the sense of constituting a reasonable first approximation, is relatively unproblematic, and it is difficult to mount a substantive criticism of the rational approach on these grounds.

Interestingly, there appears to be a significant correspondence between the mental representational structure implied by a rational analysis and those derived through the construction of psychological spaces. Nosofsky (1991) details the relationship between the Generalized Context Model and the rational model of categorisation, demonstrating, in particular, that the former may be regarded as a special case of the latter. Furthermore, Anderson (1990) argues that the rational approach to categorisation mirrors Shepard's (1987a) analysis of psychological spaces in terms of consequential regions and resultant generalisation gradients, in the sense that:

“both start with minimal, reasonable assumptions about the structure of the environment, then derive the probability that a generalization is valid given that structure, and then make the assumption that behavior will reflect this objectively derived probability” (p. 133)

In an extension of this observation, Anderson (1992) proposes that the fundamental difference between the two approaches involves Shepard's (1987a) assumption that each point in a psychological space is equally likely to represent a stimulus, reflecting a uniform distribution of stimulus probability within a given consequential region, whilst the rational approach assumes a normal distribution of stimulus probability within a given consequential region. Unfortunately, Anderson (1992, p. 423, personal communication, May 1995) reports that the effect of this difference upon the theoretically derived generalisation gradients has not been conclusively determined, and hence an empirical determination of the relative appropriateness of the two approaches has not been conducted. Whilst such discrepancies potentially exist, the application of

the quantitative formalisation employed under the rational approach to the learning of psychological space representations would appear to lack complete justification. In any case, the focus of rational models upon stimuli which assume discrete values on component dimensions makes their application to the continuously varying dimensions of psychological spaces less than straight forward.

At a general level, nevertheless, both the rational and psychological space approaches clearly have much in common, and constitute powerful Bayesian analyses of the adaptive relationship between human mental representation and the structure of the environment. In this sense, the impressive ability of rational models to emulate human cognitive processes serves to reinforce the appropriateness of the psychological space representational construct. Ultimately, however, the difficulties involved in employing the quantitative detail of the rational approach to the modelling task at hand means that other formalisations of the relationship between the mind and the world must be examined.

### 6.1.2. Psychological Essentialism

Medin and Ortony (1989) propose the theory of 'psychological essentialism' as a means of explicating the relationship between the mental representations of objects and the consequences and properties of those objects. Basically, psychological essentialism suggests that although a conceptualisation of the structure of the world based upon notions of Platonic forms - a philosophical stance known as essentialism - is inappropriate as a model of the nature of reality, humans do tend to perceive the objects in their environment as possessing 'essences' because of the predictive and explanatory advantages afforded by this representational framework. Put simply, Medin and Ortony (1989) suggest that essentialism may properly be regarded as bad ontology, but may also happen to constitute good epistemology.

Psychological essentialism does not assert that humans necessarily are able to specify the precise nature of the essence associated with a particular object, but rather suggests that "people find it natural to assume, or to act as though, concepts have essences" (Medin & Ortony, 1989, p. 184). Consequently, some considerable flexibility exists in describing the way in which mental representational essences become linked to stimuli in the external environment. In fact, psychological essentialism suggests that the abstract structures to which human mental representation ultimately takes recourse are causally linked to sensory stimulus descriptions. This establishment of a meaningful relationship between abstract mental representation and sensory stimulation allows Medin and Ortony (1989) to make the insightful observation that:

"organisms have evolved in such a way that their perceptual (and conceptual) systems are sensitive to just those kinds of similarity that lead them towards deeper and more central properties" (p. 186)

As a concrete example of this relationship, consider a natural kind such as 'tree'. It seems likely (see Rosch 1978) that if sets of outline drawings of trees are suitably averaged, the resultant outline drawings are, in general, recognisable as trees. In terms of psychological essentialism, this result suggests that the mental representational ascription of a tree essence to each instance of tree encountered through evolutionary history has resulted in the construction of perceptual machinery which has been finely tuned to view the outline drawings as a collection of physical features which constitute a tree. Sensory description which is biased in this way clearly serves the adaptive advantage of immediately providing information regarding the similarity of a stimulus to previously encountered stimuli. Indeed, to the extent that "psychological similarity is tuned to those superficial properties that are likely to be causally linked at a deeper level" (Medin & Ortony, 1989, p. 186), the sensory description approach to mental representation is, through the evolutionary adaptation of human perceptual structures, largely appropriate.

This conclusion is not incompatible with those reached in Section 2.2.2, where the primacy of sensory stimulation as a means of encountering the world, and the occasional appropriateness of physical descriptions as models of mental representation was noted. Nevertheless, particularly with regard to humans, the arguments advanced in Section 2.2.2 in relation to the frequent inappropriateness of the sensory description approach do imply that human perceptual machinery has not evolved to the point where conceptual structure can always be inferred directly from sensory experience. Whilst the composite outline of a number of trees may be recognisable as a tree, the same is not true for the composite outline of a number of pieces of furniture (again, see Rosch 1978). The perceptual amalgam of a chair, table, closet and bookcase, for instance, is extremely unlikely to resemble any discernible object, and almost certainly will not evoke the concept 'furniture'.

Observations of this type, of course, form the foundation of the notion of a 'basic' level in conceptual hierarchies (see Mervis & Rosch 1981, Rosch 1978, recall Section 3.1.5). It is interesting, therefore, to observe evidence that basic level concepts are the first learned by humans, and are assigned linguistic labels which are relatively simpler (Rosch, Mervis, Gray, Johnson & Boyes-Braem 1976). As Neisser (1987) notes: "categorisation at the basic level is categorisation by appearances" (p. 14). In terms of the evolution of mental representation, such results suggest that perceptual similarity may provide an initial basis for the representation of the external world. This basis can subsequently be manipulated to form the more abstract representational structures which provide further predictive and explanatory advantages.

This is essentially the conclusion reached by Goldstone (1994) in developing a 'bootstrapping' model of the acquisition of conceptual structure. Having identified the problems with sensory description and featural abstraction discussed in Chapter 2, by suggesting that similarity-based cognitive models:

“must successfully navigate between the Scylla of a purely perceptual basis and the Charybdis of an unconstrained set of postulated aspects” (p. 146)

Goldstone (1994) then argues that:

“There exists a continuum from low-level perceptual features to highly abstract theories. Explanatory progress occurs when concepts at more abstract levels are explained, in part, by concepts at lower levels” (p. 146)

The weight of empirical developmental evidence and theoretical explanatory power presented by Goldstone (1994) in support of this view of the evolution of conceptual structures is considerable, and indicates, in part, that sensory description has an important role in defining the relationship between the mind and the world. Clearly, however, the acceptance of the existence of abstract causal linkages underscores the fact that sensory description is incomplete as an approach to the modelling of human mental representation.

In this regard, Medin and Ortony (1989) emphasise that the theory of psychological essentialism recognises that the physical properties of stimuli are potentially superficial, and need not be the sole determinant of the mental representation of those stimuli. Rather, it is suggested that the same representation may be ascribed to a set of stimuli on the basis of some common consequence, where that consequence is causally associated with a particular representational essence. Averill’s (1993) discussion of psychological essentialism is more explicit in this regard, identifying plausible manifestations of underlying essences with regard to natural kinds, biological kinds and functional kinds. In all cases, however, the proposed nature of the causation linking an essence to its emergent property is presented in terms of intervening ‘theories’, of the type which characterise the explanation-based approach to understanding human conceptual structure (Komatsu 1992, Medin 1989, Smith 1989). Thus, all stimuli which prove to be poisonous when eaten by a particular individual become associated with a ‘poisonous’ essence, and the successive realisation of these associations promotes the individual’s mental development of a theory of poisonousness. Such a theory would, presumably, encapsulate an association of certain physical properties with poisonous stimuli, and include a belief that the nature of poison itself, through chemical or other means, is causally related to the development of these physical features in poisonous stimuli.

### **6.1.3. Psychophysical Complementarity**

Shepard (1975, 1981b, 1984, 1987b) has developed, primarily under the title ‘psychophysical complementarity’, a set of notions which are largely consonant with psychological essentialism. In particular, Shepard’s (1981b) advocacy of psychophysical complementarity incorporates the fundamental insight of psychological essentialism, that of the existence of a causal linkage between mental representation and the perception of the world:

“selective pressures of biological evolution ... have shaped, in higher organisms, a



perceptual mechanism whereby objects are represented in a way which preserves the information most essential for survival - information about the inherent properties of objects and the organism's spatial relations to them" (p. 291)

Rather than employing a formalisation of linking theories, however, psychophysical complementarity conceives of the relationship between objects and their mental representations in terms of abstract functional isomorphisms. Within this conceptual scheme, what are termed the 'inherent properties' of objects possess a structure which necessitates a complementary mental representational structure. Appropriate complementarity is not achieved, however, through a 'first-order' isomorphism (Shepard 1975), in which the components of mental representation can be canonically mapped onto components of the inherent structure of the external environment. Instead, 'second-order' isomorphisms (again, see Shepard 1975) are invoked, in which canonical *functional* relationships are established between mental representations and the world.

In describing this notion of complementarity, Shepard (1981b) employs the metaphor of a lock's functional relationship to a key.

"just as a lock has a hidden structure that is to some extent complementary to the visible contour of the key, the internal [mental representational] structure that is uniquely activated by a given object must have a structure that somehow meshes with the pattern manifested by this object" (p. 291)

The functional isomorphisms of psychophysical complementarity define a relationship between the mind and the world which appears very similar in spirit to the notion of abstract causal association which underpins psychological essentialism. As with the linking theories of psychological essentialism, the functional isomorphisms between mental representations and the external environment confer adaptive advantage. Specifically, a natural outcome of the operation of internal cognitive processes upon functionally based representations is a heightened perceptual and physical preparedness, following stimulation, on the part of an organism. A mental representation which is functionally 'tuned' to make immediate the poisonousness of a spider, for example, is ideally constructed to realise the appropriate attentional strictures required of perceptual mechanisms in gauging the danger inherent in such an experience, and also to issue the necessary motor actions based upon consequent perception. Effectively, the mental representation of the spider, and the sensory stimulation provided by the spider itself, form a 'mesh', in which cognitive conceptual structures and physical realities are coherently bound to predictive, explanatory and adaptive advantage. As Garner (1993) concludes:

"[the] complementarity idea is that the real world and the Mind can function together in a way which neither could alone ... the two must mesh so they can jointly perform a needed function" (p. 170)

#### 6.1.4. Process-Based Representation

Carlton and Shepard (1990a) note that the representational dictates of theories such as

psychological essentialism and psychophysical complementarity suggest that “it is not objects, but their transformations which are primary” (p. 129). The notion that mental representations correspond to the parameters of some form of reconstructive process, thus forming an efficient and informative abstract coding of stimuli in the external world is intuitively appealing, and has received considerable empirical and theoretical support.

For example, studies of apparent motion (see Shepard & Cooper 1982, Carlton & Shepard 1990a, 1990b), including particularly those involving mental rotation, provide strong evidence in favour of modelling the mental representation of stimuli in terms of the geometric transformations which are applied to the stimuli. Notions such as ‘representational momentum’ (see, for example, Freyd 1987) provide a complementary impetus for viewing mental representation in dynamic terms. Similarly supportive is Leyton’s (1992) conceptualisation of the processes which act on non-rigid stimuli as being mentally encoded through the symmetry properties of the geometric transformations which deform the stimuli (see also Palmer 1991, for a related set of ideas).

Vickers’ (1996, see also Vickers, Vincent & Medvedev 1996) proposed ‘Erlanger’ program for psychology further develops this general approach, in seeking to apply notions of geometrical and topological invariants to cognitive modelling. The name ‘Erlanger’ reflects the inspiration of this approach in the influential and powerful ‘Erlanger Program’ for geometry articulated by Felix Klein in 1872. Vickers (1996) identifies the foundations of the Erlanger approach in Klein’s conceptualisation that:

“different geometries can be characterised by different *groups of transformations*, and the important relationships within each geometry are defined by the *symmetries* or set of properties which remain *invariant* under the corresponding group of transformations” (p. 3)

and argues that this view has successfully been employed to systematise and extend both the fields of mathematics and physics. The insight afforded by the Erlanger approach is neatly captured by Kac and Ulam’s (1968) observation that:

“[t]he fruitfulness of this point of view stems from the fact that algebraic properties of the group of transformations that leave a certain ... structure invariant may reflect many of the properties of the structure itself” (p. 72)

Thus, in applying the Erlanger conceptualisation to psychology, Vickers (1996) proposes that:

“the single most important function carried out by the brain is to perform multiple geometric transformations on patterns of incoming sensory excitation, and all significant mental events and processes are determined by invariants associated with these transformations or to which they give rise through continued iteration” (p. 2)

This view of the relationship between the mind and the world closely accords with that advanced in Section 3.3.2 - that of the mind reflecting the world as mediated by cognition and perception, ensuring an integration of the learning of mental representational into the general

cognitive operation of human in the world - to justify the development of a radial basis function implementation of multidimensional scaling. The Erlanger approach suggests that mental representations correspond to those parameters, or seed elements, of complex perceptual and cognitive processes which allow the reconstruction of, or otherwise 'resonate' with, the sensory experiences being represented. The mental representational of a distal stimulus, under this view, constitutes a 'tuning' or 'meshing' of cognitive and perceptual processes to the associated proximal stimulus.

#### 6.1.5. Physical Constraints

A final contribution of psychological theory towards understanding the relationship between the human mind and the world emphasises the importance of the actual physical structure of the human body. Glenberg (in press) neatly encapsulates this contribution by arguing that:

“conceptualization is the encoding of patterns of possible physical interaction with a three-dimensional world. These patterns are constrained by the structure of the environment, the structure of our bodies, and memory. Thus, how we perceive and conceive of the environment is determined by the types of bodies we have” (p. 1)

Glenberg (in press) provides an analysis of a range of memory and language phenomena in terms of their dependence upon the exact nature of the human sensors and effectors through which they are realised. Lakoff (1987) places similar importance on the structure of the human body in a linguistically based analysis of conceptual structure. For example, the whole-part dichotomy which underpins mental hierarchies is viewed as being closely related to the natural subdivision of human body parts. Such approaches are not inconsistent with views such as psychological essentialism and psychophysical complementarity. Clearly, all incorporate attempts to understand the development of human mental representation in terms of existing cognitive structures and information provided by the external environment. The introduction of the structure of the human body as an additional influence on mental representation, however, constitutes a significant extension of this position.

Humphrey's (1992) consideration of the evolution of consciousness and mental representation argues for the primacy of the structure of an organism as a means of explicating the relationship between the mind and the brain. From an evolutionary standpoint, the argument is that:

“boundaries - and the physical structures that constituted them, membranes, skins - were crucial. First, they held the animal's substance in, and the rest of the world out. Second, by virtue of being at the animal's surface they formed a frontier: the frontier at which the outside world impacted upon the animal, and across which exchanges of matter and energy and information could take place” (Humphrey 1992, p. 18)

In other words, an understanding of the way in which the external world constrains mental representation must necessarily incorporate some understanding of the medium through which the

physical world interacts with the mental. In terms of accounting for the incremental development of cognitive structures across evolutionary history, Humphrey's (1992) emphasis on the physical structure of organisms affords considerable insight and is largely convincing - particularly in relation to the initial emergence of rudimentary cognitive capabilities in the most primitive organisms. At this level, Humphrey's (1992) argument is somewhat reminiscent of Goldstone's (1994, recall Section 6.1.2) 'bootstrapping' conception of the development of mental structures in humans. Despite this apparent correspondence, however, the importance attached to the physical structure of human bodies by Glenberg (in press), Lakoff (1987) and others could be questioned.

Doubts regarding the primacy of the form of the human body as an explanatory mechanism in relation to human mental structure are, perhaps, best articulated in terms of the appropriateness of the so called 'Turing Test' for artificial intelligence (Turing 1950, see also Hofstadter 1985, chap. 22). In essence, the Turing Test involves the human interrogation of a purportedly intelligent agent by means of a typed 'conversation' conducted through computer terminals. The view espoused by Turing (1950) is that if the human interrogator cannot, after prolonged interaction with the agent in this way, determine whether or not the agent is human, and the agent is in fact some type of artificial system, then that system is properly regarded as exhibiting artificial intelligence.

The Turing Test has been criticised (eg. Brooks 1991b, p. 573) for its general compliance with symbolic approaches to artificial intelligence (eg. Newell 1980, Newell & Simon 1976), and particularly with regard to the disembodied notion of intelligence it appears to encapsulate. Many aspects of intelligent behaviour would appear to be inaccessible to the interrogator of a Turing Test because of the significantly impoverished means of communication upon which the test relies. Nevertheless, intelligence does seem capable of maintaining a certain degree of abstraction from particular physical manifestations. Certainly, the view that intelligence can only be accommodated within the physical structure of a human is unjustifiably anthropocentric. Indeed, given the close coupling between mental representation, cognitive processing and intelligence asserted in Chapter 1, this thesis' subsequent modelling of mental representational structure in non-biological hardware involves the implicit assumption that intelligent information processing can be realised in a variety of computational media.

Thus, the Turing Test may be considered as occupying one end of a spectrum of intelligence testing approaches which progresses towards successively more natural forms of human interaction between the interrogator and agent. It is impossible to determine the point at which the consequent involvement of physical human structure in such a gradation ceases to allow the assessment of important aspects of intelligence, and begins to inappropriately handicap an artificial agent merely on the basis of its physical appearance. This dilemma is, in essence, a restatement of the difficulty inherent in assessing the appropriate emphasis to place on the human body's influence upon mental representation.

Perhaps a reasonable position to adopt on this issue is the following: Whilst the explication of mental structure directly in terms of body structure overstates the role human physical structure plays in conveying environmental information, the general approach adopted by Glenberg (in press), Lakoff (1987), Humphrey (1992) and others serves to emphasise that environmental constraints on mental representation arise through both the perceptual and motor interactions of humans with the world. Theories such as psychological essentialism and psychophysical complementarity sometimes appear susceptible to considering the relationship between the mind and the world solely in terms of the perceptual constraints sensation imposes upon cognition. In principle, however, both theories are completely capable of also incorporating - and, indeed, they occasionally explicitly recognise - representational constraints arising from human interaction with the world generated through movement. To the extent that an emphasis on the physical structure of the human body serves to reinforce the existence and importance of these haptic constraints it should be regarded as a valuable contribution towards an understanding of the relationship between the mind and the world.

#### 6.1.6. Conclusion

There appears to be a significant basis for viewing the impact of the external world upon human behaviour as extending to the structuring of mental representation. Moreover, the conceptualisations of the relationship between the mind and the world considered above are largely compatible with the psychological space approach to modelling mental representation advocated in Chapter 3. What psychological essentialism describes as the representational 'essence' of a stimuli seems to parallel the 'mental elements' of the rational approach, the 'inherent properties' of psychophysical complementarity, and the 'transformational invariants' or 'seed elements' of the Erlanger program. Such abstract representational essences would, in turn, appear to be appropriately modelled in abstract representational psychological spaces.

There also appears to be some broad agreement on the nature and purpose of the relationship between the world and the mental representational structures it constrains. Simon and Kaplan (1989) provide a neat summary:

“Intelligent systems are ground between the nether millstone of their physiology or hardware, which sets inner limits on their adaptation, and the upper millstone of a complex environment which places demands on them for change” (p. 38)

Essentially, therefore, the relationship between the mind and the world is one based on adaptation. The development of mental representational structures releases an organism from reliance on the continual provision of sensory information, and allows the organism to predict and comprehend the world in a significantly enhanced way. However, the precise means by which such environmental constraints on mental representation are appropriately formalised remain elusive.

The difficulties inherent in applying the rational approach to a connectionist model of the development of psychological spaces were discussed in Section 6.1.2. With regard to psychological essentialism, the explanatory flexibility achieved by considering conceptual structure as a conglomeration of theories in this way comes at the expense of the possibility of developing formal modelling mechanisms. As noted by Komatsu (1992), the means by which the requisite linking theories of an explanation-based approach such as psychological essentialism are appropriately implemented in a connectionist network, or any other formal modelling framework, are enormously difficult to determine. Psychophysical complementarity's notion of a second order isomorphism forming a functional mesh between a distal object and its mental representation is also not readily amenable to connectionist implementation. Shepard (1989) is explicitly concerned with such modelling possibilities, but, in general, provides an insightful programmatic discussion rather than proposing specific network mechanisms and structures. Similarly, Glenberg (in press), in relation to the development of models which incorporate the human bodies influence upon mental structure, states that:

“it may well be that connectionism will be the surest route to formalizing these ideas. Nonetheless, it will have to be a connectionism that differs from the sorts currently in use” (p. 19)

Thus, it is difficult to apply psychological theory directly to the development of connectionist mechanisms which internally derive indices of psychological similarity. Therefore, the most promising approach towards the construction of a workable mechanism would appear to involve examining standard connectionist practices regarding the representation of information which has its source in an external environment, and then refining and extending these techniques under the guidance of the psychological theory surveyed above.

---

## 6.2. Connectionist Internalisation Of Psychological Similarity

The ability to construct models which are embodied and situated in an environment was advanced in Chapter 1 as one of the primary cognitive modelling attractions of connectionism. It is not surprising, therefore, that the majority of connectionist cognitive models incorporate feedback which can, in some sense, be considered as modelling information provided by a simulated external world. In particular, the learning processes which operate within previously developed models are commonly driven by information regarding what may be described as the *categorical associations* and *sensory properties* of presented stimuli.

### 6.2.1. Categorical Associations And Sensory Properties

The prototypical connectionist approach to the provision of categorical associations involves a layer of output units which are constructed in one-to-one correspondence with a set of categories

to which elements of the stimulus domain may belong. The pattern of category membership of a presented stimulus across these categories is then formalised through feedback which specifies appropriate activation values for the output units, in accordance with some representational convention. The only restriction typically placed on such conventions is, in accordance with the isomorphism between environmental categories and network units, that a local representational scheme be employed - although there is some evidence (eg. Markman 1989) that relatively more minor representational details can significantly influence the capabilities of a model. In any case, a number of models described in Chapter 2 and Chapter 3 provide examples of this general category unit approach, including the disease diagnosis model, the ARTMAP mushroom model, the Neocognitron model of letter recognition, and the ALEX and ALCOVE models. Each of these models is essentially concerned with the cognitive process of categorisation, and relies upon feedback which provides correct categorical associations through activating category units corresponding to the disease categories 'burlosis' or 'terrigitis', the categories 'poisonousness' or 'edible', the letters 'a' through 'z', or whatever range of environmental categories must be accommodated by the model.

Connectionist semantic networks (eg. Rumelhart & Todd 1993, recall Section 2.3.1) employ a different approach of the representation of categorical associations. The theoretical heritage of connectionist semantic networks results in external feedback being formalised in terms of the relational properties and attributes of the stimuli. Whether or not this approach significantly extends the category unit approach depends upon the degree to which it can be approximated through the combinatorial construction of category units which combine relational, property and attribute units (recall Figure 2.10). For example, it may be the case that the environmental feedback provided by a connectionist semantic network through the dual activation of a relational 'has-a' unit and a 'beak' unit can be emulated by the creation of a lone 'has-a-beak' or, simply, 'beaked' category unit. It is also possible, however, that the semantic network approach's introduction of what effectively amounts to distributed category representation creates significant differences in the computational properties and representational abilities of the two approaches. At the very least, the representational distribution of environmental information evident in connectionist semantic networks allows for the parsimonious construction of network architectures.

A second approach towards the formalisation of the external feedback received by connectionist models involves the provision of physical or sensory information which describes the various properties of stimuli. This approach is particularly prevalent in non-psychological modelling involving the training of networks to correctly predict specified sensory properties (eg. Lawrence, Tsoi & Back 1996), but is, in principle, equally applicable in psychological contexts. The NETtalk model of speech production (Sejnowski & Rosenberg 1987), for example, employs feedback which is formalised in terms of phonemic codes and, as such, is closely related to sensory

acoustic information. A similar state of affairs is evident in the connectionist speech recognition model described by Kohonen (1988b), which establishes a phonemic space through self-organising map learning techniques, and then recognises spoken words on the basis of the trajectories they impart upon this space. Clearly, the provision of information through sensory description allows a relatively simple and effective means of modelling the external environment in terms directly adapted from the physical sciences. Furthermore, this approach is entirely compatible with the embodied view of human cognition, in which sensory boundaries between humans and the world serve as the primary mediators of environmental information.

It could reasonably be argued, however, that the lack of detail involved in connectionist models employing either categorical associations or sensory properties approach often results in an uncomfortably arbitrary correspondence between the provided feedback and the environmental events which are being modelled. In particular, simulated environments are often entirely static, in one, or both, of the following two senses.

First, the environment is often not viewed as being inherently dynamic, in that the presentation of a particular stimulus is always followed by the provision of complete and invariant feedback information. Simulated environments seldom contain any notion of the continually changing nature of the external world. As Brooks (1991b) argues, the potential for this type of deficiency in any simulated model of an environment is only circumvented through the construction of cognitive models which physically interact with the real world. Employing Brooks' (1991b, p. 583) slogan "the world is its own best model", this is precisely the approach adopted by the 'Artificial Life' sub-field of Artificial Intelligence which, in general, pursues the development of robots and other physical agents which perform elementary behavioural tasks in real world environments. Unfortunately, this research tends to eschew the consideration of cognitive representation (see Brooks 1991a) and, indeed, it is difficult to reconcile many mental representational modelling goals with the Artificial Life approach.

There is, however, no fundamental barrier preventing the employment of simulated environments which are significantly more realistic than is currently standard connectionist practice. It is possible to provide incomplete feedback with regard to the categorical associations of a stimulus on some trials, reflecting, for example, the fact that not every real world observation of a robin necessarily indicates whether or not it is a member of the category 'capable of flight'. Similarly, feedback regarding the sensory properties of a stimulus is amenable to psychophysically principled formalisations, incorporating, for example, the presence of stochastically varying sensory noise (see Rumelhart & Todd 1993, p. 12, for an example).

Secondly, simulated environments are often static in the sense that the information processing being performed by a connectionist model does not alter the feedback patterns it receives. It is in this context that a recognition of the relationship between mental representation



and the physical structure and actions of the human body, as discussed in Section 6.1.5, is particularly valuable. By viewing connectionist models as embodied and situated agents, it is possible to incorporate the belief that “inseparable from the perceived attributes of objects are the ways in which humans habitually use or interact with those objects” (Rosch 1978, p. 33). In particular, as was outlined in Chapter 1, it is possible to associate output units within a network as representing the activation of motor responses which, through an appropriately sophisticated simulation of the model’s environment, alter the subsequent feedback information provided to the model.

Therefore, whilst accepting inherent limitations in modelling interaction with the external world in terms of information regarding the categorical associations and sensory properties of stimuli, it appears reasonable to suggest that such an approach offers a basic technique for introducing environmental information to a connectionist network which learns psychological space representations.

The architectural basis of this proposed approach is schematised in Figure 6.1. In essence, the presentation of a stimulus to the model at the stimulus input layer results in the prediction of the categorical associations and sensory properties of that stimulus across the output layer. With reference to Figures 3.7 and 3.8, it is clear that the ability to make such prediction partly involves the learning of weighted connections between the exemplar and output layers, as is evident in the ALEX and ALCOVE models. It is from these connection weights that indices of inter-stimulus psychological similarity may be derived.

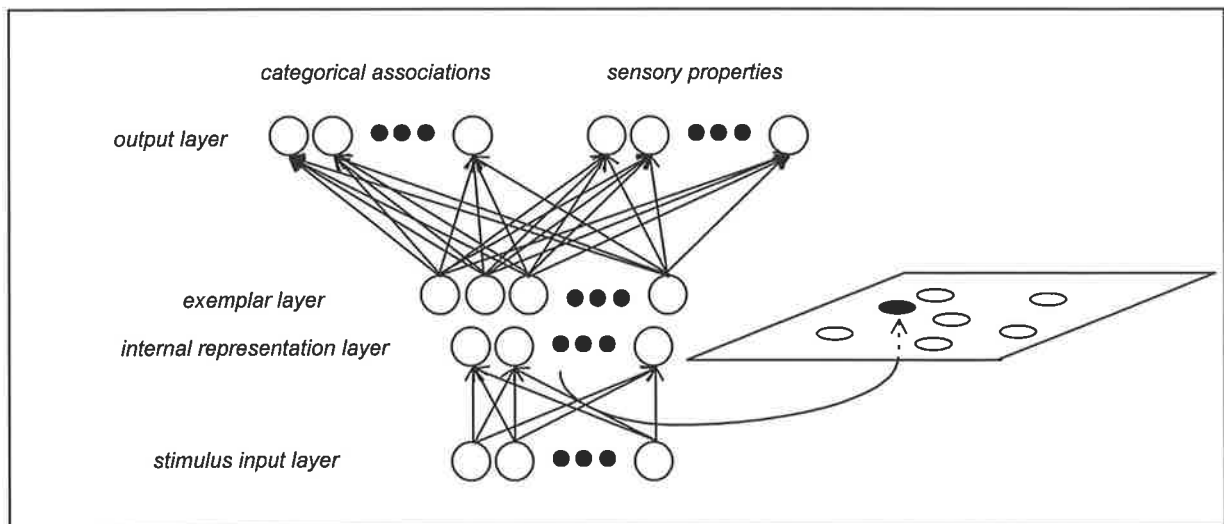


Figure 6.1. The proposed approach to environmentally constraining the learning of mental representation.

Given appropriate measures of psychological similarity, the model architecture shown in Figure 6.1 also indicates the way in which mental representational structures can be developed. In particular, the stimulus input, internal representation, and exemplar layers correspond precisely to those of the connectionist multidimensional scaling network developed in Chapter 4, as depicted in Figure 4.1. Thus, the learning of the psychological space representational locations of the stimulus

set involves the adjustment of the connection weights linking the stimulus input and internal representation layers.

### 6.2.2. A First Approximation Of Psychological Similarity

Having described a formalisation of the external environment's influence upon a connectionist model, it is appropriate to examine the plausibility of this formalisation in terms of the general theoretical development of the relationship between the mind and the world presented in Section 6.1. Fortunately, the provision of the categorical associations and sensory properties of stimuli does appear to be in broad agreement with the established theoretical conception of this relationship. The sensory information provided by a stimulus, and the higher-level cognitive concepts to which that stimulus belongs, seem to constitute precisely the type of information humans acquire from interaction with the environment which serves to constrain the adaptive development of their mental representational structures.

Moreover, considerable specific precedent may be found with regard to the derivation of indices of psychological similarity from categorical and sensory information. Wallach's (1958) survey of various measures of psychological similarity, for example, argues for the partial appropriateness of definitions cast in terms of "common environmental properties" (p. 105), and the cognitive tendency to "assign items to a common category" (p. 106). Clearly, these definitions closely correspond to the notions of sensory properties and categorical association, respectively, as developed above. It is, therefore, reassuring to note that Gregson's (1975) critique of Wallach's (1958) survey identifies the possibility of psychological similarity being defined in terms of common category assignment as a "serious theoretical contender" (p. 14).

Further support for the notion of measuring psychological similarity in terms of categorical associations is provided by so-called 'classificatory theories' of similarity (eg. Sjöberg & Thorslund 1979). Essentially, such theories, supported by empirically observed relationships between classificatory responses and similarity ratings (eg. Handel 1967), suggest that the psychological similarity of two stimuli is a function of the classificatory properties, or categorical associations, of those stimuli. A set of closely related assumptions underpin additive clustering techniques (Shepard 1980, Shepard & Arable 1979, Tenenbaum 1996) for describing conceptual structures. These techniques assume that:

"stimuli are represented as members of salient subsets (presumably corresponding to natural classes or features in the world) and similarity is treated as a weighted sum of common and distinctive subsets" (Tenenbaum 1996)

In other words, as with classificatory theories, the psychological similarity of two stimuli is viewed as arising from their common categorical associations.

Finally, the notion of consequential regions (recall Section 3.1.2) provides considerable support for seeking to derive psychological similarity from both the categorical associations and



sensory properties of stimuli. Not only are psychological space representations viewed as being adaptively constructed subject to environmental constraints (Shepard 1987a, 1987b, 1994), but consequential regions can readily be identified with both categorical and sensory information. The postulated learning processes which establish the boundaries of consequential regions operate through the revision of gradients of generalisation based upon information provided by an external environment (Shepard 1994, Shepard & Kannappan 1991, Shepard & Tenenbaum 1991). This information may be purely sensory in nature, as in the determination of a 'poisonousness' consequential region through monitoring of sensory experiences of gustatory pain. It seems equally plausible, however, to suggest that environmental information may be of a more abstract categorical nature, characterising the type of feedback which a parent might provide to a child, or, indeed, which any informed source might provide to a cognitive agent. Under this scheme, for example, a student's consequential region corresponding to the natural kind 'communist' is refined through information regarding the categorical associations of various historical figures in relation to the concept 'communist', as provided by lecturers, textbooks, and so on.

Thus, the process by which the boundaries of consequential regions are modified may be related to the development of increasingly accurate measures of psychological similarity, as information regarding the categorical associations and sensory properties of stimuli is accrued (cf. Shepard 1965). Therefore, given the assumption which underpins the model developed in Chapter 4 that human mental representation is appropriately modelled by psychological spaces, the introduction of environmental information to constrain the learning of such representational structures seems appropriately formalised through the dual notions of categorical association and sensory properties. In particular, it would appear to be reasonable to generate indices of inter-stimulus psychological similarity from internalised measures of categorical association and sensory properties.

### 6.2.3. An Anticipatory Rejoinder

At this point, it should be acknowledged that the proposed technique for internally deriving the measures of psychological similarity appears to risk the "vacuity and circularity" described by Rips (1989). The definition of psychological similarity in terms of the sensory properties of stimuli, and the categories to which they belong, seems to necessitate the model assuming the information it is designed to learn. As Rips (1989) argues:

"if you explain why people classify bats as mammals by saying that bats are similar to other mammals, you cannot simultaneously explain that similarity by invoking shared predicates such as *is a mammal*" (p. 51).

Indeed, it could be maintained that the specification of the range of potential categorical associations of a stimulus set amounts to precisely the type of pre-abstraction which was criticised in Section 2.1 as a means of modelling mental representation within connectionist networks.

Ultimately, this line of reasoning might conclude that the proposed approach to developing internal measures of psychological similarity simply involves the transferal of inappropriate pre-abstraction from the input to the output units of a model.

As plausible as this criticism may seem, it fails to recognise a fundamental difference between the proposed scheme for environmentally constraining mental representation, and the featural approach to stimulus representation. Put simply, the important distinction to be drawn concerns the subject of the pre-abstraction. In the case of the featural approach, it is the mental representations themselves which are being modelled through the articulation of sets of psychological features. In the categorical association and sensory property approach, it is the structure of the world which is being abstracted in preparedness for presentation to a model. In the first case, the model assumes the mental representations it should seek to learn. In the second case, the model is being provided with the information it requires to learn.

As discussed in Section 6.2.1, models which are not physically embodied and able to interact with the real world must operate in some type of simulated environment. The specification of relevant categorical associations and sensory properties constitutes a series of ontological assumptions which render tenable such simulation. For example, the reliance of connectionist semantic networks upon categorical associations, in the form the various properties, qualities, actions, and so on, of a stimulus set constitutes an entirely justifiable approach to modelling the world. As Rumelhart and Todd (1993) argue:

“the network essentially ends up reflecting the structure of the world (as we human parse it, since we make the training sets)” (p. 19)

The fact that, in some cases, the range of categorical associations formalised within such models may appear dangerously close to providing the mental representational structure the model is to acquire simply reflects an introspective accuracy in the pre-abstraction of the modeller.

It is probably fair to suggest, however, that the modelling of the world solely in terms of a fixed set of categorical associations and sensory properties of stimuli is inherently limited. Indeed, the intricacies of the external environment, and their impact upon mental representation, would appear to be somewhat trivialised by the proposed approach. In this sense, the approach constitutes a first tentative formalisation of a method for modelling the acquisition of connectionist mental representation. Whilst the principles of environmentally constraining mental representation, and modelling the world through the categorical associations and sensory properties of stimuli are well founded, there remains considerable scope for the development of more sophisticated and realistic formalisations. In this regard, the addition of noise to such sensory information and the incorporation of dynamic aspects in the model of the environment, as discussed earlier, constitute ways in which an environment might be more realistically simulated.

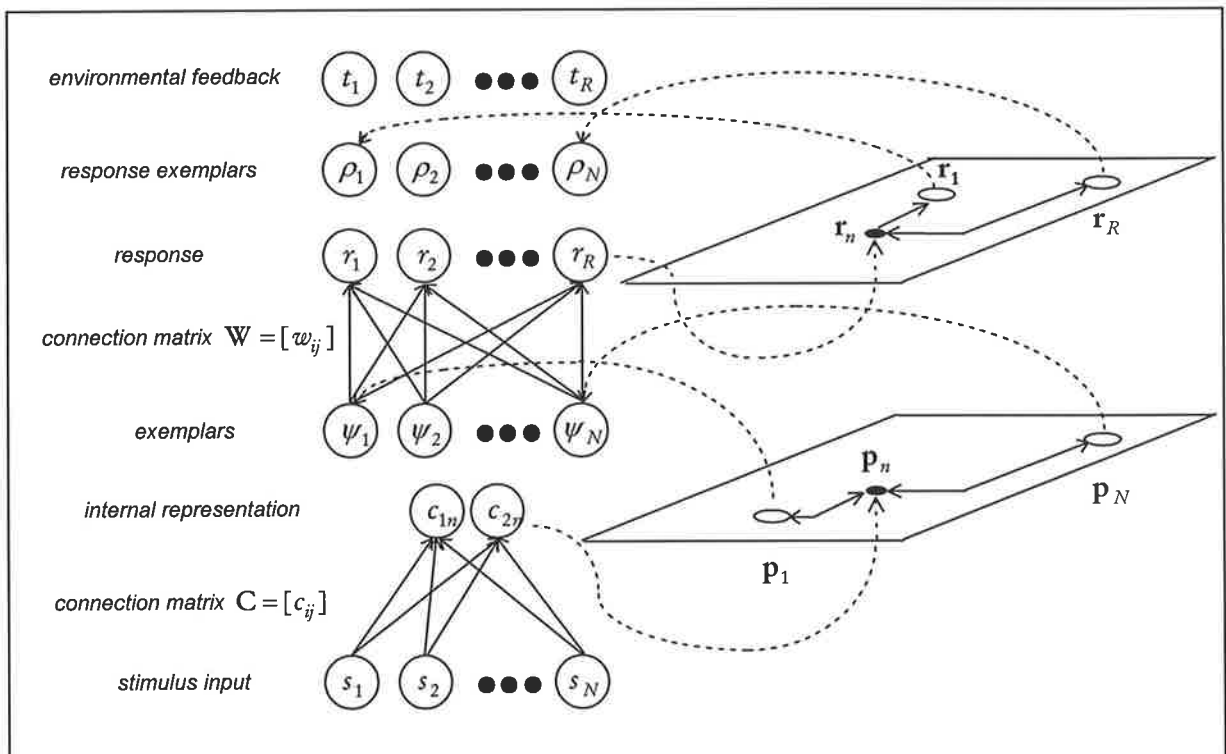
Nevertheless, even static formalisations of the sensory properties and categorical associations

of a set of stimuli, as employed in previous connectionist modelling, provide a means of allowing an external environment to constrain the representation learned by a model. This information allows a simulated world - however primitive - to dictate and impinge upon the simulated mind modelled by a connectionist network. Thus, incorporating the learning of sensory and categorical information into the model developed in Chapter 4 offers the promise of developing a connectionist model of the learning of environmentally constrained mental representation.

## Chapter 7: A Model Of The Learning Of Mental Representation

This chapter describes a connectionist model which develops psychological space internal representations using environmental information concerning the categorical associations and sensory properties of a set of stimuli. The model is founded upon the connectionist multidimensional scaling model developed in Chapter 4, appending mechanisms which allow the model to categorise stimuli and predict their sensory properties. The acquisition of knowledge required for such prediction enables the model to derive internal indices of inter-stimulus psychological similarity, which replace the externally derived measures required by the multidimensional scaling model.

The architecture and nomenclature of the model are shown in Figure 7.1. The stimulus input, internal representation and exemplar layers are directly adapted from the multidimensional scaling model. The exemplar layer is now connected to a response layer, through a second connection weight matrix  $\mathbb{W}$ . The  $R$  units in the response and environmental feedback layers are constructed in one-to-one correspondence with the range of possible environmental categorical associations and sensory properties of the stimulus set. The  $N$  units in the response exemplar layer, however, correspond to the elements of the stimulus set, and share a radial basis function linkage with the response layer.



*Figure 7.1. The architecture and nomenclature of the connectionist model of the environmentally constrained learning of mental representation.*

Once again, the model's description is readily divided into processing and learning phases. In the processing phase, a stimulus is presented, causing the generation of predicted categorical associations and sensory properties at the response layer. Following the provision of

environmentally derived information at the environmental feedback layer, the learning phase serves to adjust both the model's predictive abilities and its internal representation of the presented stimulus.

---

## 7.1. Processing Phase

The processing phase consists of six main stages: stimulus presentation, internal representation determination, similarity calculation, response generation, target similarity derivation, and environmental feedback provision. The first three of these stages are identical to those of the multidimensional scaling model, as detailed in Sections 4.1.1, 4.1.2 and 4.1.3, so it is only necessary to describe to the final three stages.

### 7.1.1. Response Generation

The pattern of activation values across the exemplar layer, which measures the current psychological similarity of each member of the stimulus set to the presented stimulus, cause the generation of activation values across the response layer through the connection weight matrix  $\mathbb{W} = [w_{ij}]$  with  $R$  rows and  $N$  columns. These response units correspond to the possible categorical associations and sensory properties of the stimulus set, with the  $i$ th unit's activation value, denoted  $r_i$ , being given by:

$$r_i = \sum_{j=1}^N \psi_j w_{ij} \tag{7.1}$$

### 7.1.2. Target Similarity Derivation

The radial basis function linkage shared by the response and response exemplar layers requires the units in the response exemplar layer to maintain a position within a 'response space'. As depicted in Figure 7.1, this is an  $R$  dimensional space, with coordinate axes corresponding to the units in the response layer. The response exemplar units assume positions within this space in accordance with the currently predicted categorical associations and sensory properties of the stimulus set. Thus, the location of the  $n$ th response exemplar unit, corresponding to the currently presented stimulus, is updated to be given by the  $R$  dimensional vector  $\mathbf{r}_n = (r_1, r_2, \dots, r_R)$ , reflecting the categorical and sensory expectations of the model generated at the response layer.

Effectively, the positioning of the response exemplar units in response space represents the model's internalisation of categorical and sensory information relating to the stimulus set, as provided by the environment. The spatial distribution of the response exemplar units directly reflects the constraining effect of the external environment upon the model. Following the discussion of Chapter 6, therefore, it is from the structure of the response space, particularly the

locations of the response exemplar units, that indices of stimulus similarity may be derived.

More specifically, the activation values assumed by the response exemplar units, in accordance with their radial basis function linkage to the response layer, should provide target measures of psychological similarity between the currently presented stimulus and the other members of the stimulus set. In effect, the activation values of the response exemplar units fulfil the role played by the external feedback layer of the multidimensional scaling model described in Chapter 4.

Clearly, a formal specification of the means by which the response exemplar units acquire appropriate activation values involves describing both the distance metric operating within response space, and the form of the radial basis function applied to distances measured according to this metric. With regard to the distance metric, it seems reasonable to argue that the categorical associations and sensory properties which define the component dimensions of response space are appropriately regarded as independent, in essentially the same sense that separable stimuli have independent component dimensions. Indeed, the fact that certain categories and properties are selected to simulate an external environment suggests that they are capable of unitary and coherent description and, consequently, may be considered in one-to-one association with units in a connectionist layer. Such injective mappings, in turn, are indicative of the representational independence of the categories and properties which underpin the response space. There is, therefore, considerable justification in assuming the operation of the City-block distance metric in response space.

The determination of an appropriate basis function, however, is markedly more problematic. In essence, the basis function must calculate an appropriate target value of the similarity between a member of the stimulus set and the presented stimulus, given the distance between the response exemplar units of the two stimuli. That is, having been provided with a distance gauging the similarity of the model's expectations with regards the categorical associations and sensory properties of two stimuli, the basis function is required to generate a measure of inter-stimulus psychological similarity.

A number of demands can reasonably be made of this basis function. Most fundamentally, the implication of Chapter 6 in general, and Section 6.2.2 in particular, is that the derived psychological similarity should monotonically decrease as distance in response space increases. Essentially, such a stricture implements the notion of mental representation being environmentally constrained by requiring that derived measures of psychological similarity reflect the shared categorical associations and proximal sensory properties of stimuli. In particular, the similarity of two stimuli - in the sense of being expected to belong to the same category, or providing the same sensory information - affords a degree of psychological similarity to be accommodated by the mental representational structure.



In addition to implementing target psychological similarity as a monotonically decreasing function of distance in response space, it also seems reasonable to suggest that the basis function should have a restricted range between, say, the values zero, indicating complete psychological dissimilarity, and one, indicating complete similarity. Therefore, denoting the response space radial basis function as  $g(D)$ , where  $D$  is the provided distance, the following restrictions may be imposed:

$$\begin{aligned}
 g: [0, \infty) &\mapsto [0, 1] \\
 g(0) &= 1 \\
 \lim_{D \rightarrow \infty} g(D) &= 0 \\
 g'(D) &\leq 0
 \end{aligned}
 \tag{7.2}$$

There are, however, an enormous variety of functional forms which satisfy these requirements, even when only continuous and otherwise ‘well-behaved’ functions are considered. Unfortunately, it is difficult to find convincing psychological evidence, of either an empirical or theoretical origin, which places additional constraints upon the response space basis function.

Some guidance, however, is provided by studies of response generalisation, which examine the extent to which a stimulus elicits responses other than that which has been learned or is otherwise intended. As noted by Shepard (1957), observed confusions in a paired-associates learning task may derive from stimulus generalisation (or confusion), response generalisation (or confusion), or a combination of both. Predominantly, empirical studies of paired-associate learning have focused upon stimulus generalisation, by allowing a range of stimuli but restricting response behaviours to a limited, discrete, and highly discriminable set of alternatives. It is possible, however, to alter this approach towards the consideration of response generalisation by restricting the range of presented stimuli to a discrete and highly discernible set, and allowing a continuum of response behaviour.

For example, Noble and Bahrck (1956) employ a pressure control as a means of producing identification responses for stimuli which signify target response pressures. Similarly, a paired associate learning task reported by Shepard (1958a, Experiment II) involves responses being made by the physical placement of an electric probe upon a linear array of contacts. Both of these studies provide empirical confirmation of the suggestion that response generalisation monotonically decreases as ‘response distance’ - as measured by the difference between target and actual pressure, or the distance between the correct and actual contact point - increases. Shepard’s (1958a) analysis extends to the construction of a representational space in which the various responses are represented by points in one dimension. This representational structure is largely equivalent to the actual physical configuration of the contacts, except that the contacts at each end of the array are relatively more distant from their neighbours (see Shepard 1958a, Figure 5). Thus, in this case, there is a close relationship between physical response descriptions and psychological response

descriptions, and it appears plausible that such a finding might generalise across other forms of responding.

Shepard's (1958a) further finding that the obtained response generalisation measures are well modelled by an exponential decay function over the derived psychological response configuration, is, therefore, highly suggestive of a general form for the basis function in response space. Additional impetus is provided by Shepard's (1958b) observation that the trace model of generalisation mentioned in Section 3.1.2, although primarily developed in terms of stimulus generalisation, is equally applicable to response generalisation. That is, the conclusion of the trace model that stimulus generalisation is given by an exponential decay function of distance in psychological (stimulus) space can be taken to imply that generalisation in response space also exponentially decays. Finally, Shepard (1957) arrives at the same conclusion, evidently on the basis of inferred symmetries or likenesses between stimulus and response spaces. Thus, what evidence is available suggests that the generation of measures of psychological similarity in terms of response differences might be appropriately generated using an exponentially decaying radial basis function.

Certainly, exponential decay functions satisfy the conditions imposed in Equations 7.2. Indeed, it might be argued that exponential decay functions canonically meet these requirements. In any case, the model developed here assumes that the radial basis function operating in response space is given by:

$$\rho_i = \exp(-\kappa|\mathbf{r}_i - \mathbf{r}_n|_1) \quad (7.3)$$

where  $\kappa$  is a non-negative parameter controlling the information properties of response space, and is discussed in detail in Section 7.3. The activation values generated across the response exemplar layer in this way constitute the model's internalised measure of the psychological similarity between the presented stimulus and every other member of the stimulus set.

### 7.1.3. Environmental Feedback Provision

The interpretation of both the environmental feedback and the response layer's activation values relies upon the adoption of some form of representational convention. Effectively, this convention reflects the assumptions made in simulating the environment within which the model is situated. With regard to categorical associations, for example, activation values of +1 might be taken as indicating that a stimulus is a member of a particular category, whilst activation values of -1 signify non-membership. With regard to sensory properties, less abstract coding schemes are readily applicable, such as the representation of a line segment through an activation value proportional to the segment's length.

Following the determination of such a representational code, the environmental feedback received by the model following the presentation of a stimulus is formalised by setting the

activation values of the environmental feedback layer to appropriate values. As discussed in Section 6.2.1, the provision of environmental information in this way allows for the possibility of incomplete, erroneous, or noise-perturbed feedback being received by the model.

---

## 7.2. Learning Phase

The learning phase consists of two main stages. First, the connection weight matrix  $W$  is adjusted according to an *external error* which measures the discrepancy between the categorical associations and sensory properties predicted by the model and those provided by the environment. Secondly, the connection weight matrix  $C$  is adjusted according to an *internal error* which measures the discrepancy between the current internal representational structure, and that required to realise a psychological space representation..

### 7.2.1. External Error

The external error,  $EE$ , is defined to be (proportional) to the sum of the squared difference between the predicted categorical association and sensory property values given across the response layer, and the values provided by the environmental feedback layer, as follows:

$$EE = \frac{1}{2} \sum_{j=1}^R (t_j - r_j)^2 \quad (7.4)$$

The learning rule derived from the external error measure acts upon each weight in the connection weight matrix  $W$ , and again adopts the gradient descent approach to optimisation. Thus, the learning rule takes the form:

$$w_{ij}^{new} = w_{ij}^{old} - \lambda_w \frac{\partial EE}{\partial w_{ij}} \quad (7.5)$$

where  $\lambda_w$  is a learning rate parameter.

Specifically, the required partial derivative, with reference to Equations 7.1 and 7.4, is calculated as follows:

$$\begin{aligned} \frac{\partial EE}{\partial w_{ij}} &= \frac{\partial}{\partial x_{ij}} \frac{1}{2} \sum_{j=1}^R (t_j - r_j)^2 \\ &= - \sum_{j=1}^R (t_j - r_j) \frac{\partial r_j}{\partial w_{ij}} \\ &= - \sum_{j=1}^R (t_j - r_j) \psi_j \end{aligned} \quad (7.6)$$

which results in the learning rule:

$$w_{ij}^{new} = w_{ij}^{old} + \lambda_w \sum_{j=1}^N (t_j - r_j) \psi_j \quad (7.7)$$

### 7.2.2. Internal Error

The internal error,  $IE$ , fulfils the role played by the total error measure,  $E^{tot}$ , in the multidimensional scaling model, and seeks to generate a psychological space internal representation through the adjustment of the connection weight matrix  $C$ . Thus, the internal error incorporates an internal similarity error component,  $IE^{sim}$ , and an internal dimensional error component,  $IE^{dim}$ , as follows (cf. Equation 4.13):

$$IE = IE^{sim} + IE^{dim} \quad (7.8)$$

The similarity error component measures the discrepancy between the pattern of inter-stimulus psychological similarities in the current internal representational structure, given by the activation values across the exemplar layer, and the target similarity values internally generated at the response exemplar layer. In particular, the error measure is proportional to the sum of the squared difference between these two sets of values (cf. Equation 4.6):

$$IE^{sim} = \frac{1}{2} \sum_{j=1}^N (\rho_j - \psi_j)^2 \quad (7.9)$$

The dimensional error component of the internal error measure is identical to the dimensional error developed for the multidimensional scaling model. Thus, with reference to Equations 4.8, 4.11 and 4.12, the dimensional error component may be defined simply as follows:

$$IE^{dim} = E^{dim} \quad (7.10)$$

It seems reasonable to consider the internal derived measures of inter-stimulus similarity generated across the response exemplar layer as being momentarily stable, at least during the processing and learning phases which follow the presentation of one particular stimulus. In this sense, the effect of modifications of the connection weight matrix  $C$  upon response exemplar activations is justifiably ignored. This simplification, in turn, allows the partial derivative of the dimensional error component of the internal error measure to be equated with that of the multidimensional scaling model's dimensional error, as derived in Equation 4.17. Thus, the learning rule associated with the internal error measure is closely related to the learning rule employed in the multidimensional scaling model, and is given by (refer Equations 4.15, 4.16, 4.17

and 4.18):

$$\begin{aligned}
c_{in}^{new} &= c_{in}^{old} - \lambda_c \frac{\partial E}{\partial c_{in}} \\
&= c_{in}^{old} - \lambda_c \sum_{j=1}^N (\rho_j - \psi_j) \psi_j \|\mathbf{p}_n, \mathbf{p}_j\|_r^{1-r} |c_{in} - c_{ij}|^{r-1} \text{sgn}(c_{in} - c_{ij}) \\
&\quad - \exp(-\beta m_i) (c_{in} - \frac{1}{N} \sum_{k=1}^N c_{ik})
\end{aligned} \tag{7.11}$$

where  $r$ , as before, defines the Minkowski distance metric assumed to operate within psychological space.

---

### 7.3. Model Construction And Parameter Setting

Much of the construction and interpretation of the model developed in this chapter parallels that of the multidimensional scaling model described in Chapter 4. In particular, the units in stimulus input, exemplar and response exemplar layers are created in one-to-one correspondence with the members of a pre-determined stimulus set, and the initial dimensionality of psychological space is over-estimated by the number of units in the internal representation layer. The foundation of the current model in the multidimensional scaling model enables the interpretation of the learned psychological space representation to be undertaken in the same way in both models. Additional architectural concerns, however, arise from the current model's simulation of the environment. Specifically, units in the response and environmental feedback layers are established in one-to-one correspondence with appropriate categorical associations and sensory properties of the stimuli.

In terms of parameter values, many of the conclusions reached in relation to the multidimensional scaling can also validly be applied to the current model. The interaction of the two learning rules does, however, suggest that some care might be taken in setting the learning rate parameters  $\lambda_c$  and  $\lambda_w$ . As the internal learning rule alters the psychological space location of the exemplar units, the connection weights previously set by the external learning rule become inaccurate. It may, therefore, be prudent to employ a value of  $\lambda_w$  which is relatively larger than  $\lambda_c$  to ensure that the model's environmental predictions maintain sufficient accuracy to give rise to meaningful target similarity values. Beyond this consideration, however, both of the learning rate parameters are appropriately set to small but essentially arbitrary values.

Whilst the dimensionality reduction parameter remains capable of significantly influencing the learned internal representational structure, there does not seem to be any reason for believing that the observed insensitivity of the multidimensional scaling model to this parameter, as demonstrated in Section 5.1.1, will not prevail in the current model. Of more concern is the

information parameter,  $\kappa$ , introduced in Equation 7.3, which is also capable of significantly affecting the model's behaviour. Accordingly, the following section develops a principled means for setting the value of this parameter.

### 7.3.1. Basis For Setting The Information Parameter

The information parameter influences the distribution of the internal indices of psychological similarity maintained at the response exemplar layer. Recall from Equation 7.3 that  $\kappa$  parameterises the exponential decay basis function employed in generating similarity measures from response space distances. As was discussed in Section 4.1.3 in relation to psychological spaces, setting  $\kappa$  to near-zero values will result in all similarity indices being close to one, whilst large  $\kappa$  values will result in near-zero similarity indices, with the sole exception of self-similarities which will maintain the appropriate value of one.

Such considerations suggest that the setting of the  $\kappa$  parameter might best be accomplished through considering the response space in general, and the target similarities in particular, as the internally maintained *information source* from which the psychological space representation is developed. Clearly, if  $\kappa$  were set to zero, the resultant target similarities, all assuming the value one, would provide little information regarding the appropriate psychological space representational structure. Similarly, large values of  $\kappa$ , causing one target similarity to be one, with the rest being zero, also provide relatively little information to the remainder of the model. In information theoretic terms, both of these parameter values restrict the flow of information from the environmental feedback layer to the internal representational layer. The response space is, in essence, being limited in its ability to convey the information necessary to develop appropriately detailed psychological space internal representations.

There is some considerable precedent regarding the analysis of human cognitive processes and representations in terms of information theory (see Garner 1962, chap. 1, for an overview), and this practice maintains contemporary popularity. As broad recent examples, consider Linsker's (1988) use of information maximisation principles in modelling the development of orientation specific visual cells, Myung's (1994) employment of entropy measures in analysing various categorisation models, and Corter and Gluck's (1992) explanation of conceptual structure and basic levels in terms of the relative informativeness of candidate representational structures. The general success of these, and other, analyses bears testimony to the potential insights gained from the application of information theory in general, and the entropy measure in particular, to a psychological context. Thus, the approach to setting the information parameter  $\kappa$  which is pursued here views the response space as an information carrying channel, and seeks to maximise an entropic measure of the information represented across the response exemplar layer. That is,  $\kappa$  assumes values producing patterns of target similarities which, in terms of an entropy measure, are

maximally informative.

Clearly, the formalisation of this approach requires the development of a quantitative understanding of the way in which target similarities are influenced by changes in the value of  $\kappa$ . Since a general means of parameter setting is sought, it is not possible to specify in advance the location of the response exemplar units within the response space. Thus, it is necessary to measure the probability distribution of target similarities given minimal assumptions regarding the position of the response exemplar units.

### 7.3.2. Distribution Of Target Similarities

The derivation of the required probability distributions uses the fact that response space is a coordinate space of dimensionality  $R$  operating under the City-block distance metric, and that target similarities are generated from the position of the response exemplar units in the way defined by Equation 7.3. It is further assumed that, first, every point in response space constitutes an equally likely location for a response exemplar unit and, secondly, that the probability of generalisation between the categorical associations or sensory properties of any two stimuli, whilst possibly negligibly small, is non-zero.

A direct consequence of this second assumption is that every response exemplar unit in response space lies within a finite distance of every other unit. Therefore, there exists a minimal positive number  $v$ , such that response exemplar units in response space differ by at most  $v$  with respect to their positions on each of the coordinate axes. Geometrically,  $v$  is the value of the length of the side of the smallest possible (hyper)cube which encompasses all of the response exemplar units in response space. Given this characterisation, it is clear that only the values of  $R$ ,  $v$  and  $\kappa$  affect the distribution of the patterns of derived target similarities.

Since the target similarities are calculated as an exponential decay function of various distances in response space, it is necessary to determine the probability distribution of these distances. Considering first the case of a one-dimensional response space, the cumulative density function of the distance between two points drawn from a uniform distribution on a line segment of length  $v$  is required. This can be found by allowing a point  $p$  to move from 0 to  $v$  along the segment, and integrating the probability that a second point,  $p'$ , chosen independently of  $p$ , will fall within a distance  $d$  of  $p$ .

As depicted in Figure 7.3, there are three cases to be considered. The first applies when  $p$  is between  $d$  and  $v-d$ . In this case, the  $p'$  can lie anywhere in the marked region between  $p-d$  and  $p+d$ , of length  $2d$ , and be within  $d$  of  $p$ . The second case applies when  $p$  is between 0 and  $d$ . Here,  $p'$  must lie between 0 and  $p+d$  to be within  $d$  of  $p$ , a region of length  $p+d$ . The final case applies when  $p$  is between  $v-d$  and  $p'$  must lie between  $p-d$  and  $v$ , a region of length  $d+(v-p)$ . Given that  $p$  and  $p'$  take each point on the line segment with equal probability, and that the locations of  $p$  and  $p'$  are

independently chosen, the required cumulative density function is derived as follows:

$$\begin{aligned}
 D_1(d) &= \int_d^{v-d} \frac{dp}{v} \times \frac{2d}{v} + \int_0^d \frac{dp}{v} \times \frac{p+d}{v} + \int_{v-d}^v \frac{dp}{v} \times \frac{d+(v-d)}{v} \\
 &= \frac{2}{v^2} [dp]_d^{v-d} + \frac{1}{v^2} \left[ \frac{2dp+p^2}{2} \right]_0^d + \frac{1}{v^2} \left[ \frac{2dp+2vp-p^2}{2} \right]_{v-d}^v \\
 &= -\left(\frac{d}{v}\right)^2 + 2\left(\frac{d}{v}\right)
 \end{aligned} \tag{7.12}$$

This result can be used to derive the cumulative distribution function for distance in higher dimensional response spaces. Because of the operation of the City-block metric is that the distance between two points located in a  $R$  dimensional space is simply the sum, across all  $R$  coordinate axes, of the distance between the two points on each individual dimension.

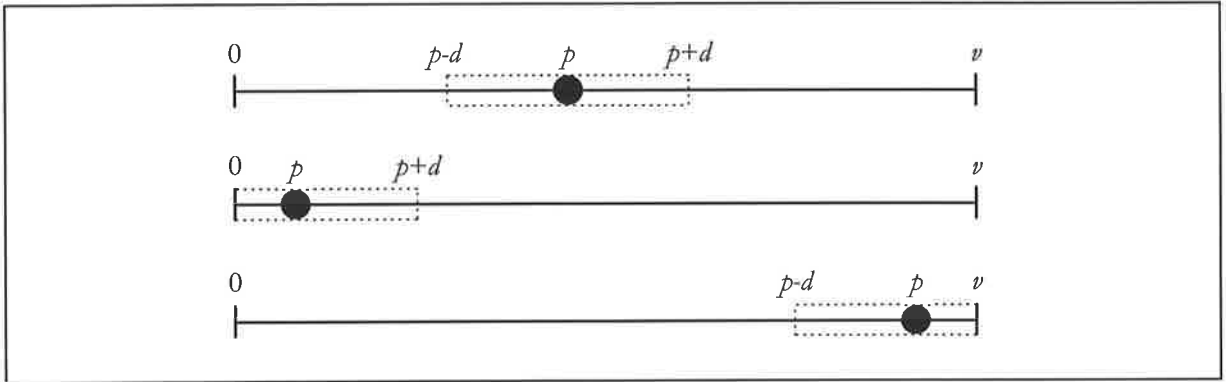


Figure 7.3. Three cases in developing the distribution of distance between two points on a line segment.

For example, consider two points,  $\mathbf{p}$  and  $\mathbf{p}'$ , in a two-dimensional response space. Let the difference between these two points on one dimension be  $u$ . Then the probability that the total distance between  $\mathbf{p}$  and  $\mathbf{p}'$  is less than  $d$  can be found by allowing  $u$  to move from its minimum of 0 to its maximum of  $v$ , and integrating the probability that the sum of the distance on the first dimension and the distance on the second dimension is less than  $d$ .

There are now two cases to be considered. In the first case, where  $0 \leq d \leq v$ ,  $u$  must lie between 0 and  $d$ , and the second dimension's distance must be less than  $d-u$ . Thus, for  $0 \leq d \leq v$ :

$$\begin{aligned}
 D_{2,1}(d) &= \int_0^d D_1'(u) \cdot D_1(d-u) \cdot du \\
 &= \int_0^d \left(-\frac{2u}{v^2} + \frac{2}{v}\right) \times \left(-\frac{(d-u)^2}{v^2} + \frac{2(d-u)}{v}\right) \cdot du \\
 &= \frac{1}{6} \left(\frac{d}{v}\right)^4 - \frac{4}{3} \left(\frac{d}{v}\right)^3 + 2\left(\frac{d}{v}\right)^2
 \end{aligned} \tag{7.13}$$

In the second case, where  $v \leq d \leq 2v$ , either  $u$  lies between  $d-v$  and  $v$  and the second dimension's distance is less than  $d-u$ , or  $u$  is less than  $d-v$  and the second dimension's distance can assume any value. Thus, for  $v \leq d \leq 2v$ :



$$\begin{aligned}
D_{2,2}(d) &= D_1(d-v) + \int_{d-v}^d D_1'(u) \cdot D_1(d-u) \cdot du & (7.14) \\
&= \frac{2(d-v)}{v} - \frac{(d-v)^2}{v^2} + \int_{d-v}^v \left(-\frac{2u}{v^2} + \frac{2}{v}\right) \times \left(-\frac{(d-u)^2}{v^2} + \frac{2(d-u)}{v}\right) \cdot du \\
&= -\frac{1}{6} \left(\frac{d}{v}\right)^4 + \frac{4}{3} \left(\frac{d}{v}\right)^3 - 4 \left(\frac{d}{v}\right)^2 + \frac{16}{3} \left(\frac{d}{v}\right) - \frac{5}{3}
\end{aligned}$$

The results of Equations 7.13 and 7.14 constitute a piece-wise definition of the distribution of distance in a two dimensional response space.

A general means of deriving the distance distribution in a response space of arbitrary dimensionality is evident from the construction of the two dimensional distribution. For a response space of dimensionality  $R$ , the distance distribution is comprised of  $R$  functions, each of which can be expressed in terms of the distribution of distance in  $R-1$  and one dimensional spaces, as follows:

$$\begin{aligned}
D_{R,1}(d) &= \int_0^d D_{R-1}'(u) \cdot D_1(d-u) \cdot du & (7.15) \\
D_{R,2}(d), \dots, D_{R,R-1}(d) &= D_{R-1}(d-v) + \int_{d-v}^d D_{R-1}'(u) \cdot D_1(d-u) \cdot du \\
D_{R,R}(d) &= D_{R-1}(d-v) + \int_{d-v}^{(R-1)v} D_{R-1}'(u) \cdot D_1(d-u) \cdot du
\end{aligned}$$

where  $D_{R,x}(d)$  applies on the closed interval  $[(x-1)v, xv]$ . Since the distribution function  $D_{R-1}(d)$  will be similarly comprised of  $R-1$  functions, Equations 7.15 can be given more precisely as follows:

$$\begin{aligned}
D_{R,1}(d) &= \int_0^d D_{R-1,1}'(u) \cdot D_1(d-u) \cdot du & (7.16) \\
D_{R,x}(d) &= D_{R-1,x-1}(d-v) + \int_{d-v}^{(x-1)v} D_{R-1,x-1}'(u) \cdot D_1(d-u) \cdot du \\
&\quad + \int_{(x-1)v}^d D_{R-1,x}'(u) \cdot D_1(d-u) \cdot du \\
D_{R,R}(d) &= D_{R-1,R-1}(d-v) + \int_{d-v}^{(R-1)v} D_{R-1,R-1}'(u) \cdot D_1(d-u) \cdot du
\end{aligned}$$

where  $x = 2, 3, \dots, R-1$ . In this way, the cumulative density function of distance in an arbitrary dimensional response can be built up recursively from the distribution of distance on a line segment, as given in Equation 7.12. It is possible to show that, despite their piece-wise definition, these functions are both continuous and differentiable on the relevant interval  $[0, Rv]$ .

The distribution of target similarities represented by the activation values across the response exemplar layer can be derived from the distance distributions. With reference to Equation 7.3, the cumulative density function of target similarities in a  $R$  dimensional response space is given by:

$$A_R(a) = p(\text{target similarity in } R \text{ dimensions} \leq a) \quad (7.17)$$

$$\begin{aligned}
&= p(\exp(-\kappa \cdot \text{distance in } R \text{ dimensions}) \leq a) \\
&= p(-\kappa \cdot \text{distance in } R \text{ dimensions} \leq \ln a) \\
&= p(\text{distance in } R \text{ dimensions} \geq \frac{-\ln a}{\kappa}) \\
&= 1 - p(\text{distance in } R \text{ dimensions} \leq \frac{-\ln a}{\kappa}) \\
&= 1 - D_R\left(\frac{-\ln a}{\kappa}\right)
\end{aligned}$$

Clearly, the cumulative distribution function for target similarity will be piece-wise comprised of  $R$  functions, following the piece-wise definition of distance distribution. Specifically:

$$A_{R,x}(a) = 1 - D_{R,x}\left(\frac{-\ln a}{\kappa}\right) \quad (7.18)$$

where  $x = 1, 2, \dots, R$ ,  $A_{R,x}(a)$  applies on the interval  $[\exp\{-\kappa \cdot xv\}, \exp\{-\kappa \cdot (x-1)v\}]$ , and  $A_R(a)$  as a whole applies to the interval  $[\exp\{-\kappa \cdot Rv\}, 1]$ . For example, the distance distributions given in Equations 7.12, 7.13 and 7.14 give rise to the following cumulative distribution functions for target similarities in one and two dimensional response spaces:

$$A_1(a) = \left(\frac{\ln a}{\kappa v}\right)^2 + 2\left(\frac{\ln a}{\kappa v}\right) + 1 \quad (7.19)$$

$$A_{2,1}(a) = -\frac{1}{6}\left(\frac{\ln a}{\kappa v}\right)^4 - \frac{4}{3}\left(\frac{\ln a}{\kappa v}\right)^3 - 2\left(\frac{\ln a}{\kappa v}\right)^2 + 1$$

$$A_{2,2}(a) = \frac{1}{6}\left(\frac{\ln a}{\kappa v}\right)^4 + \frac{4}{3}\left(\frac{\ln a}{\kappa v}\right)^3 + 4\left(\frac{\ln a}{\kappa v}\right)^2 + \frac{16}{3}\left(\frac{\ln a}{\kappa v}\right) - \frac{2}{3}$$

The continuity and differentiability of these piece-wise defined target similarity distributions follows from the continuity and differentiability of the distance distributions, and the method of construction detailed in Equation 7.16.

### 7.3.3. Maximising Entropy

The cumulative probability distribution functions of target similarity given in Equations 7.19 indicate that all such functions can be expressed as polynomials in the indeterminate  $\frac{\ln a}{\kappa v}$ . More specifically, it is clear that, for a response space of given dimensionality  $R$ , the distribution of target similarities is completely characterised by the single value  $\kappa v$  - the product of the information parameter and the maximum spread of the response exemplar units. Whilst the value of  $v$  cannot reasonably be pre-determined by the model, and, indeed, is potentially subject to variation it can, nevertheless, be measured at any stage during the model's operation from the various positions of the response exemplar units in response space. Thus, the setting of  $\kappa$  should be interpreted in terms of ensuring that the product  $\kappa v$  maintains an appropriate value.

Following earlier discussion, an appropriate value for  $\kappa v$  is assumed to be one which

maximises the analog interpretation of the entropy measure (Shannon & Weaver 1949), which, for a random variable  $X$  with a probability density function  $p(X)$  is given by:

$$H = - \int_{-\infty}^{\infty} p(X) \ln(p(X)) \cdot dX \quad (7.20)$$

For a one dimensional response space, the target similarities have the probability density function (refer Equation 7.19):

$$A'_1(a) = \frac{2}{a} \left( \frac{-\ln a}{(\kappa v)^2} + \frac{1}{\kappa v} \right) \quad (7.21)$$

allowing the definition of an entropy measure:

$$\begin{aligned} H(\kappa v) &= - \int_{\exp(-\kappa v)}^1 A'_1(a) \cdot \ln(A'_1(a)) \cdot da \\ &= \ln(\kappa v) - \frac{\kappa v}{3} + \frac{1}{2} - \ln(2) \end{aligned} \quad (7.22)$$

which, in turn, is maximised when  $\kappa v = 3$ . Thus, when  $R = 1$ , the information parameter  $\kappa$  is appropriately set to the value  $\frac{3}{v}$ . Importantly, as is evident from the graphical depiction of Equation 7.22 shown in Figure 7.4, the entropy measure is relatively large for a considerable range of  $\kappa v$  values surrounding 3. Accordingly, it seems reasonable to suggest that the information providing capabilities of a one-dimensional response space would not be significantly impeded by the setting of  $\kappa$  to a value which resulted in a  $\kappa v$  value somewhat different from 3. This state of affairs is highly desirable, given the potential variation in the value of  $v$  as response exemplar units are moved within response space to accommodate information newly received from the environment.

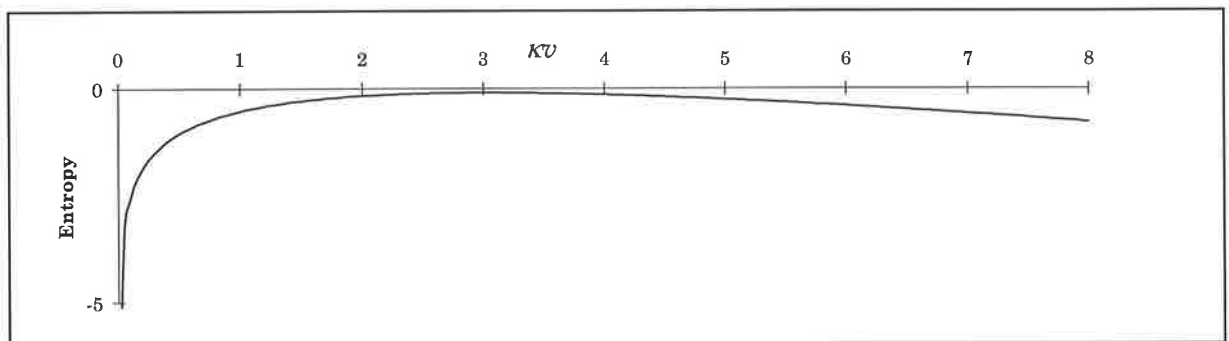


Figure 7.4. The entropy measure of target similarities in a one-dimensional response space as a function of  $\kappa v$ .

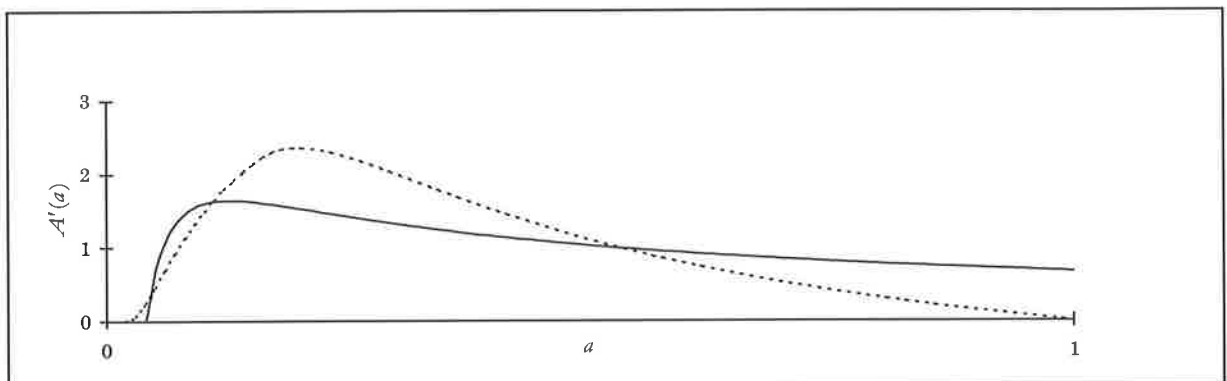
It is possible, in principle, following the method detailed in Section 7.3.2, to derive probability distributions for target similarities in a response space of any dimensionality, and then, as outlined in this section, use these distributions to define an entropy measure which can be maximised to give appropriate values for the information parameter. Unfortunately the analytic evaluation of the entropy measure undertaken for the one-dimensional case in Equation 7.22

proves to be considerably more elusive in higher dimensional cases. The general difficulty of the necessary integrations, coupled with the piece-wise definition of the probability distribution functions involved, renders the manual derivation infeasible, and specialised software for performing symbolic mathematics seems incapable of generating plausible (ie. non-complex!) solutions. Furthermore, numerical calculations of the entropy measure appear particularly volatile, presumably because of the large numbers involved in evaluating the limiting behaviour of the logarithmic function, and do not provide reliable approximations.

Therefore, in order to develop some understanding of appropriate parameter values for the information parameter in multidimensional response spaces, recourse is taken to two less principled, but computationally tractable approaches, which remain based upon the notion of maximising the information carrying capacity of response space. The first of these approaches involves finding the value of  $\kappa v$  which most closely aligns the target similarity probability density of a multidimensional response space with the probability density which is known to be optimal in the one-dimensional case. That is, by considering the distributions given in Equation 7.19 as functions of both  $a$  and  $\kappa v$ , a  $\kappa v$  value is sought which minimises the difference between the target similarity distribution in a response space with  $R$  dimensions, and the distribution  $A_1'(a,3)$ , as measured by:

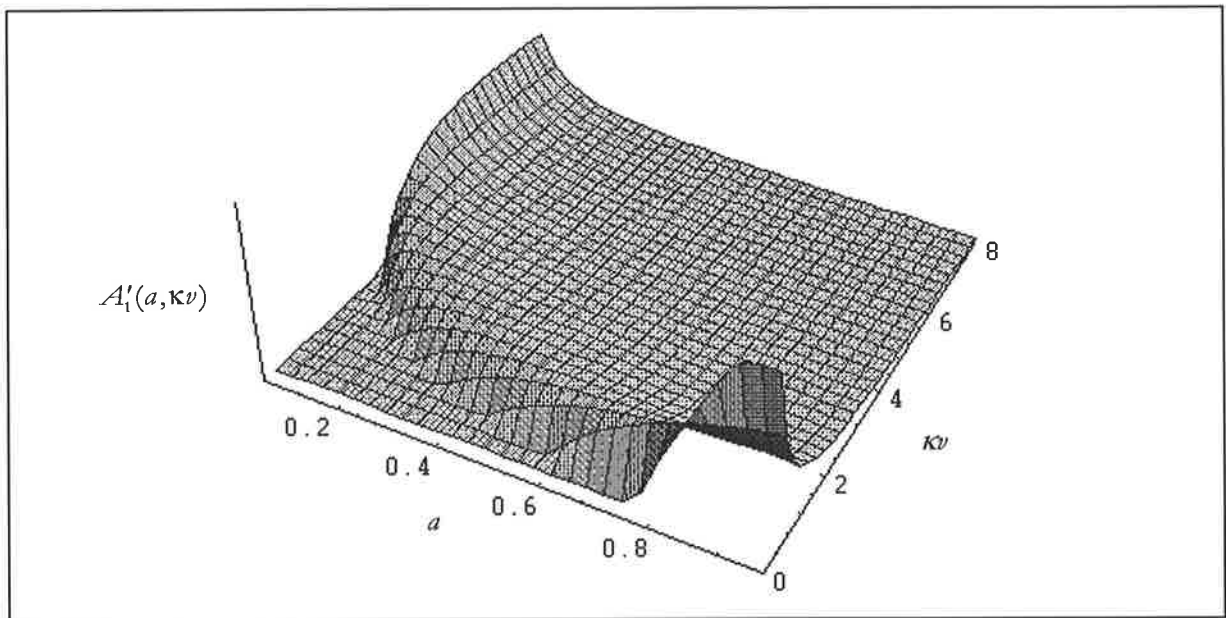
$$M(\kappa v) = \int_0^1 (A_R'(a, \kappa v) - A_1'(a, 3))^2 \cdot da \quad (7.23)$$

where both target similarity distributions take the value zero across the interval of their domains for which they are not otherwise defined. Numerical analysis of Equation 7.23 indicates that  $\kappa v$  values of approximately 1.78 and 1.27 minimise  $M$  for  $R = 2$  and  $R = 3$  respectively. Figure 7.5 shows the optimal one-dimensional distribution,  $A_1'(a,3)$  and the best fitting two-dimensional distribution,  $A_2'(a,1.78)$  as solid and broken lines respectively. Unfortunately, in response spaces with four or more dimensions, the numerical integration techniques employed again appear to become too unstable to provide reliable estimates.



*Figure 7.5. The optimal target similarity probability density function in a one-dimensional response space, and the best fitting two-dimensional distribution, shown, respectively, by solid and dotted lines.*

The second approach towards finding appropriate information parameter settings in multidimensional response spaces hinges upon a recognition that, of all probability densities with bounded domains, it is the uniform distribution which maximises the entropy measure given in Equation 7.20<sup>7</sup>. Indeed, the observed insensitivity of the entropy measure to variations of  $\kappa\nu$  in the one dimensional case, as shown in Figure 7.4 above, may be interpreted in terms of the relative uniformity of the target similarity probability density functions across these values. Figure 7.6 provides a graphical representation of this state of affairs, showing the family of similarity distributions realised in the one-dimensional case by varying the value of  $\kappa\nu$ . Between the extremes of small  $\kappa\nu$  values, in which only large similarity indices are generated, and large  $\kappa\nu$  values, in which the similarities are almost all small, there lies a significant region in which most similarity values occur with some non-negligible probability density. Allowing for a wide range of target similarity values, all of which are approximately equally likely, ensures, ultimately, that a response space contains enough information to oversee the development of a non-degenerate psychological space representation.



*Figure 7.6. The family of probability density functions of target similarities in a one-dimensional response space, across different values of  $\kappa\nu$ .*

Thus, it seems reasonable to seek  $\kappa\nu$  values which, rather than directly maximising entropy, are most similar to the uniform distribution, according to a difference analogous to that defined by Equation 7.23. More specifically, in a response space with  $R$  dimensions, the minimum of the following function is found:

$$\hat{M}(\kappa\nu) = \int_0^1 (A'_R(a, \kappa\nu) - 1)^2 \cdot da \quad (7.24)$$

The application of numerical techniques to Equation 7.24 yields  $\kappa\nu$  values of approximately

<sup>7</sup> This can be established, for example, using Lagrangian multipliers.

2.69, 1.40, 0.95, and 0.72 for one, two, three and four dimensional response spaces, respectively. The discrepancy between the known optimal value of 3 for a one-dimensional response space, and the value of 2.69 derived by this method is an immediate corollary of the fact that the probability density  $A_1'(3)$  is not uniform. The one and two-dimensional target similarity distributions which best fit the uniform distribution are shown in Figure 7.7.

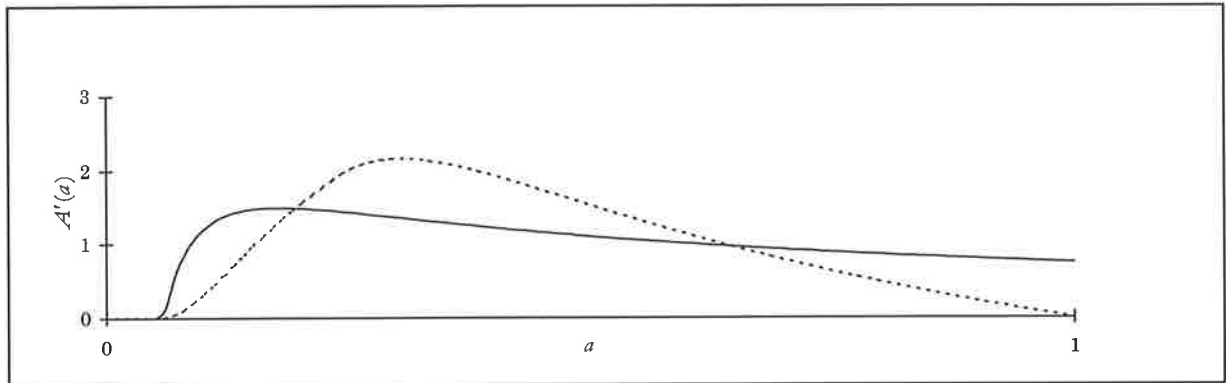


Figure 7.7. The most uniform target similarity probability density function for one and two-dimensional response spaces, shown, respectively, in solid and dotted lines.

Both of these approaches to finding values for the  $\kappa$  parameter which maximise the entropic flow of information through the model are appropriately regarded as ad hoc, and neither could be seen as being particularly accurate in its own right. Nevertheless, the preceding analysis, viewed as a whole, suggests a pattern of results which accords well with intuition, and provides something approximating a principled means for setting the information parameter. By setting  $\kappa$  in such a way as to maintain a  $\kappa v$  value of 3 for a one-dimensional response space, the pattern of target similarities are capable of being sufficiently varied to allow the construction of an appropriate psychological space.

The desired  $\kappa v$  value in multidimensional response spaces is more difficult to quantify, but its general range is more consistently identified. As the dimensionality of response space increases, the  $\kappa v$  value monotonically decreases in a negatively accelerated manner. By making essentially arbitrary assumptions about the functional form corresponding to this change in  $\kappa v$  values, a fitted curve could be generated which specified an appropriate  $\kappa v$  value as a function of the number of response space dimensions,  $R$ . This specification would merely serve to ensure that  $\kappa$  was set in such a way as to avoid the extreme distributions of target similarity associated with overly small or large  $\kappa v$  values (recall Figures 7.4 and 7.6). Such specifications can, in light of the results presented above, also be made by hand, and made explicit in the construction of a model. Whilst awaiting a technique to employ the principle of entropy maximisation in a quantitatively precise way, the second approach seems to represent more accurately the achievements of the analysis presented above, and is adopted in the demonstrations, evaluations, and extensions of the model described in following chapters.

At this point, it is worth noting that the normalisation of the axes of the configuration examined by Hubert et. al. (1992, refer Section 5.2.2) was undertaken for essentially the same reasons which dictate the adaptive setting of the information parameter. The similarity matrix generated by the application of the exponential decay function to the original distances in Hubert et. al. (1992, Figure 1a) is not sufficiently informative - in the sense that most values are near zero - to enable the connectionist multidimensional scaling model to recover the known configuration. The rescaling of the axes depicted in Figure 5.14, whilst not changing the configuration, does alter the derived similarity matrix given in Table 5.3, to one which contains a greater range of target similarity values. In effect, both the normalisation and the adjustment of the information parameter perform the same role: that of making the entropy of an information source, whether it is a similarity matrix or a response space, sufficiently large so as to allow a psychological space representation to be derived.

---

## 7.4. Overview Of The Mental Representation Learning Model

Before conducting a quantitative evaluation of the model of mental representation learning through simulation, however, it would seem appropriate to reflect upon its relationship to the previously suggested models on which it is partly based, and to discuss the model in terms of the more general psychological considerations which have guided its construction.

### 7.4.1. Relationship To Connectionist Semantic Networks And The ALEX Model

The way in which the mental representation model incorporates environmental information is motivated, as was discussed in Section 6.2, from previously developed connectionist models. In particular, Section 6.2.1 identified similarities with connectionist semantic networks, as described in Section 2.3.1, and the ALEX model, described in Section 3.2.4. Therefore, having now fully developed the model, it would seem appropriate to examine these relationships in greater detail.

Both connectionist semantic networks and the mental representation model effectively learn a set of input/output pairings which relate members of stimulus set to their various environmental properties. Most generally, connectionist semantic networks are trained to learn the properties, qualities and actions of a stimulus, as well as the relationship of that stimulus to other stimuli (recall Figure 2.10). Essentially the same environmental information is formalised within the current model in terms of the sensory properties and categorical associations of each of the stimuli. Both models acquire the same information regarding the stimulus set they must learn to internally represent.

The fundamental difference between the two models, as anticipated in Section 2.3.3, is that the current model places psychologically motivated constraints upon the internal representations it develops. Connectionist semantic networks, in seeking the derivation of 'mental' structures, rely

solely on the representational economy promoted by the inclusion of internal ‘bottleneck’ layers. The representations derived at the internal representation layer of the mental representation model, however, strictly adhere to the representational dictates of the psychological space theory. In this way, the derived internal representational structure is subjected to constraints other than those demanded by the learning of input/output pairings. These additional constraints are particularly strong, and effectively posit *psychological* grounds for preferring one mediating internal representation above all others. In particular, the internal representation of the stimuli must meet the similarity requirements of the Universal Law of Generalization in a space of minimal dimensionality.

The difference between the two models is readily cast in terms of the learning rules they employ. The backpropagation-based learning rule used by connectionist semantic networks effectively amounts to an extension of the external error learning rule of the current model and, therefore, solely acts to allow the prediction of a stimulus’ environmental properties. The mental representation model, however, restricts the use of the external error learning rule to refining the associations between the *internal* representation of stimuli and their environmental properties. A separate learning rule, defined in relation to the internal error measure, acts to develop these internal representations, and it is this learning rule which enforces both the similarity and dimensionality dictates of the psychological space theory. Interestingly, the incorporation of a bottleneck layer, perhaps the primary representational strength of connectionist semantic networks, arises naturally in the current model, since the adoption of the principle of dimensionality reduction results in the internal representational layer generally containing significantly fewer units than the stimulus input layer.

Overall, however, the current model does not merely match the mental representational claims of connectionist semantic networks. Through strict and focused adherence to a psychologically principled theory of internal representational structure, the current model presents significant grounds for being viewed as a model of the learning of mental representation.

In contrast, the similarity between the current model and the ALEX model is founded upon their shared adoption of psychological space representations. The architectural similarities which follow from this adoption, already alluded to in Section 6.2.1, are evident from an examination of Figure 7.8. Both models incorporate a layer which corresponds to the psychological space representation of a presented stimulus, and employ a radial basis function architecture to implement the Universal Law of Generalization. In addition, both models adopt an exemplar approach to the modelling of conceptual structure, with units corresponding to each element of the stimulus set being placed in the psychological space. Finally, both models are virtually identical with regard to their association of exemplar units with output responses through a set of connection weights, and the operation of a learning rule which modifies these associations



following externally provided feedback.

The fundamental difference between the current model and the ALEX model, as is also evident from Figure 7.8, relates to the way in which stimuli are presented to the model. The ALEX model is exposed to a stimulus through the setting of input activation values which correspond to the psychological space representation of that stimulus, as previously derived through some external means. The mental representation learning model, in contrast, learns to internally represent a stimulus in terms of an appropriate psychological space position. As was emphasised in Section 4.1.1, the current model receives only a nominal indication of the presence of a particular stimulus in a manner entirely devoid of structural or representational information. The inclusion of comprehensive connection weights between the stimulus input and internal representation layers, coupled with the addition of a second radial basis function structure at the response layer, allows the model to derive a psychological space representational structure.

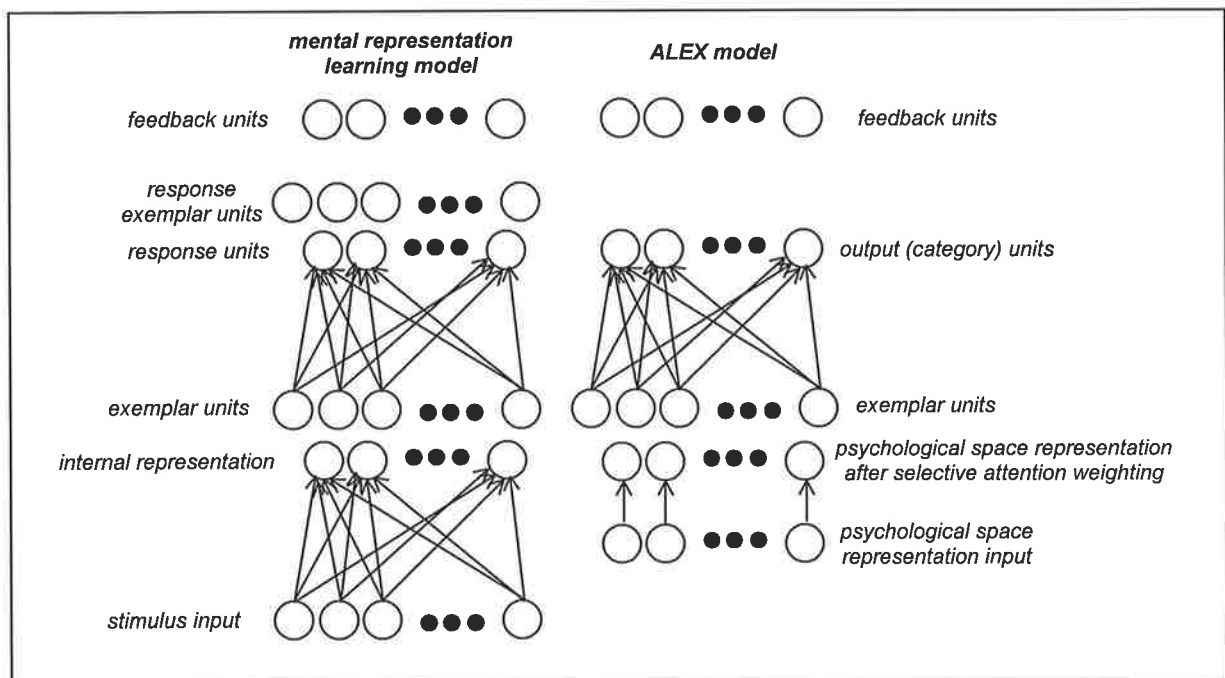


Figure 7.8. Comparison of the current model, shown on the left, and the ALEX model, shown on the right.

This difference is largely a reflection of the cognitive task domains to which the ALEX model is applied. As noted in Section 3.2.4, the ALEX model is a model of human category learning which attempts to explain and predict the way in which a given set of stimuli are, under supervision, identified as belonging to a given set of categories. In this limited situation it may be reasonable to assume that mental representations of the stimuli already exist. The mental representation learning model, however, seeks to formalise a way in which such mental representations might be learned. Indeed, given these disparate goals, the fact that the architecture and operation of the two models have much in common suggests that the impressive categorisation performance of the ALEX model, also detailed in Section 3.2.4, might also be displayed by the mental representation learning model. This is highly desirable, since the psychological space

representations developed by the model rely, in part, on the effective learning of the categorical associations of the stimulus set.

In fact, there are some grounds for suggesting that the mental representation learning model might better accommodate human categorisation phenomena in some circumstances. Because of their pre-determination, the psychological space representations employed by the ALEX model remain fixed during the operation of the model. Once again, this may be a reasonable assumption in relation to many categorisation tasks; but it is also possible to envisage a stimulus domain in which the psychological space representational structure is appropriately modified during the course of the task. Repeating the multidimensional scaling analysis to maintain the appropriateness of the ALEX model's representational structure each time such an alteration occurs would clearly be unreasonable in terms of plausibly modelling human cognition. Equally inappropriate would be the incorporation of previously developed 'off-line' techniques (Miyano & Inukai 1982) for incrementally adjusting multidimensional scaling solutions. The current model's integration of categorisation performance and internal representation development, in contrast, seems particularly suited to situations in which stimuli do not have fixed psychological space representations.

It is worth noting, however, that the mental representation learning model does not incorporate the selective attention mechanisms which significantly contribute to the performance of the ALEX model. The architectural relationships detailed in Figure 7.8 suggest that the accommodation of these abilities within the current model is feasible. In particular, if the internal representation layer of the current model were aligned with the psychological space representational input layer of the ALEX model, the inclusion of an additional layer would allow the 'primary' psychological space locations developed by the current model to be altered by the effects of selective attention.

#### **7.4.2. Relationship To Piagetian Learning Principles**

In broader psychological terms, the learning processes adopted by the mental representation model may be likened to Piagetian notions of human cognitive development. Piaget (1970, see Hilgard & Bower 1975 for an overview) noted the impact on human learning of biological maturation, experience of the physical environment, and experience of the social environment. The mental representation model does not substantially address issues of biological development or maturation, since it is intended to model the acquisition of mental representational structure in normal adult humans possessing cognitive structures which are, in the evolutionary sense, contemporary. The environmental information by which the model's learning is constrained, however, corresponds precisely to those listed by Piaget. Being provided with the sensory properties of the stimuli which are encountered constitutes experience of the physical

environment, whilst categorical associations often encode social conventions, such as identifying which members of a stimulus set are correctly regarded as 'valuable'. Given that Piaget (1970) discusses these three learning principles in the context that they are both fundamental and common to classical theories of human development, the adherence of the mental representation model to the experiential principles reinforces the appropriateness of the formalisation of environmental information developed in Section 6.2.

Despite perceiving these sources of learning as being necessary, however, Piaget (1970) argues that they are insufficient and consequently proposes a fourth learning process of 'equilibration'. In essence, the development of the notion of equilibration constitutes a theoretical attempt to specify the way in which environmental information sources are integrated within an existing conceptual structure. In Piagetian terms, equilibration describes the continuing cognitive adjustment which arises from attempting to reconcile two competing learning processes termed 'assimilation' and 'accommodation'. Assimilation refers to the process whereby newly encountered environmental information is incorporated into current conceptual structures, whilst accommodation refers to the modification of these conceptual structures in accordance with environmental experience. Assimilation, therefore, involves a direct reaction to current experience through the incremental adjustment of mental structures designed to enhance their predictive and explanatory accuracy. Importantly, these adjustments are made in the context of existing conceptual structures, emphasising the cognitive utility of interpreting the present in terms of the past. Effectively, assimilation is the fine tuning of an established mental world view to currently available environmental information. Accommodation, in contrast, involves a direct modification of underlying mental representational structures, typically necessitated by more fundamental inconsistencies between cognitive predictions and environmental feedback.

The learning principle of equilibration articulated by Piaget (1970), in seeking to balance the representational effects of these assimilatory and accommodatory processes, is entirely consistent with the relationship between the mind and the world espoused in Chapter 6. The process of accommodation insists the mind reflect the world, whilst assimilation ensures that, given such a representational foundation, cognition can serve to allow the adaptive prediction and explanation of the encompassing environment.

Furthermore, the learning processes which operate within the mental representation model bear striking similarities to those of assimilation and accommodation. The learning rule based on the external error measure, which serves to modify the associative connections between mental representations and the predicted environmental properties of stimuli, would seem to correspond to a process of assimilation. This learning rule does not alter the psychological space representational structure but, directly on the basis of current environmental feedback, incrementally adjusts the cognitive predictions made by the model. The learning rule based on the

internal error measure, in contrast, modifies the internal representational structure of the model in a process reminiscent of accommodation.

The concurrent action of the external and internal learning rules within the mental representation model, therefore, would appear to correspond closely to the equilibration process summarised by Hilgard and Bower (1975) as:

“An adjustive process ... needed to fit external reality into an existing structure (assimilation), and to modify that structure while this is taking place (accommodation)” (p. 323)

Clearly, these two processes are formalised by the two learning rules operating within the model. The internal learning rule acts to reposition the psychological space representations of the stimulus set to ‘accommodate’ the target similarities derived in the response space, and these target similarities are, in turn, generated from knowledge regarding the external world which is ‘assimilated’ in the associative weights maintained by the external learning rule. In this way, the learning of mental representation is realised as an outcome of the cognitive interaction of the model with its environment.

## Chapter 8: Demonstrations Of The Mental Representation Learning Model

This chapter demonstrates and evaluates the model of the learning of mental representation developed in Chapter 7. First, two models which receive environmental feedback in the form of sensory information are examined. Secondly, three models which rely solely on categorical associations in learning psychological space representations are described. Finally, the ability of the model to adapt its internal representational structure to a dynamically changing environment is explored.







---

### 8.1. Sensory Properties

#### 8.1.1. The Bug Model

As a first example of the mental representation model's ability to learn psychological space representations on the basis of sensory information, an environment was constructed consisting of six 'bugs' which the model could encounter. Each of the different types of bugs possessed characteristic thermal and auditory properties, as detailed in Table 8.1. Therefore, the bug model consisted of two units in the output and environmental feedback layers, corresponding to these sensory properties, and maintained six units in the stimulus input, exemplar, and response exemplar layers in accordance with the number of bugs. Once again, six units were placed in the internal representation layer. Learning rate parameters of  $\lambda_c = 0.1$  and  $\lambda_w = 0.1$  were employed, and the dimensionality reduction parameter,  $\beta$ , was set to 10. The information parameter  $\kappa$  was altered during the model's operation to maintain a target  $\kappa v$  value of 1.5, and the operation of the City-block distance metric in psychological space was assumed.

Table 8.1. Sensory Properties of Six Bugs

						
Thermal	+2	+1	-1	+2	+1	-1
Auditory	+1	+1	+1	-1	-1	-1

The coding of the sensory properties evident in Table 8.1 is particularly simple, and may be interpreted literally. For example, the first and fourth bugs can be considered to impart the greatest heat upon the model's 'sensory receptors', whereas the third and sixth bugs are relatively cold. Similarly, whilst the first three bugs emit a noise upon presentation to the model, the remaining three bugs are mute.

The addition of noise to these sensory properties, as discussed in Section 6.2.1, is important in the sense of contributing to a realistic simulation of the model's sensory experience of the

environment. Unfortunately, however, the model's assumption of an exponential decay generalisation gradient is inappropriate, at least in principle, in noisy environments. Recall from Section 3.1.5 that Ennis' (1988a, 1988b, 1992) re-derivation of the Universal Law of Generalisation for stochastically varying stimulus points in psychological space suggests that a Gaussian radial basis function should be employed. The model's reformulation in these terms appears straightforward, and is a worthwhile topic for further research. Meanwhile, however, it seems sensible to examine the ability of the model to operate in noisy environments provided the strength, or variation, of the incorporated noise is not too great.

Therefore, the sensory properties encoded in Table 8.1 were subjected to noise before being made available as environmental feedback to the model. Specifically, noise sampled from a Gaussian distribution with zero mean and a variance of 0.1 was independently added to each unit in the environmental feedback layer. The effect of this noise upon the auditory and thermal sensory information received by the model is shown in Figure 8.1.

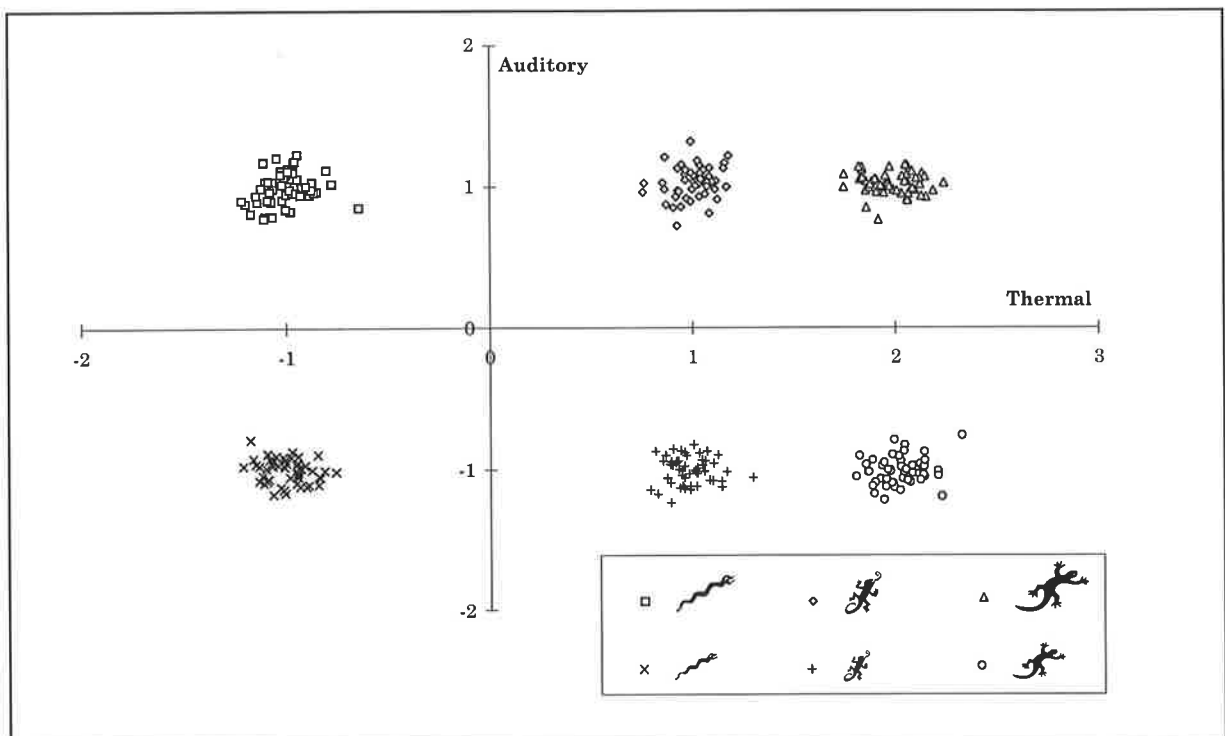


Figure 8.1. The nature of the sensory information received by the bug model. 50 samples for each bug type are shown.

The pattern of change of the external error measure across 3,000 trials is shown in Figure 8.2. After approximately 1,200 trials, the model is able to make reasonably accurate predictions of the sensory properties of each of the six bug types. Whilst the external error measure may become negligibly small, careful inspection indicates that it does not consistently achieve a value of zero. This is because predictions being made by the model correspond to the sensory values given in Table 8.1, meaning that the stochastic variation of the sensory feedback causes small errors. Nevertheless, the model's predictions are appropriate in the sense that they identify the mean of the distribution of sensory properties for each bug type, and constitute the best estimate of the

property values possible within the noisy environment. An important feature of the external error learning rule is its ability to minimise the effect of noise in this way, particularly when relatively small  $\lambda_w$  values are employed. At a similar stage, the model develops a representational structure which satisfies the internally derived indices of psychological similarity, as evidenced by the negligible value of the internal similarity error, shown in Figure 8.3. Furthermore, the stepwise decrease of the internal dimensional error indicates that this representational structure is being accommodated within a space of reduced dimensionality.

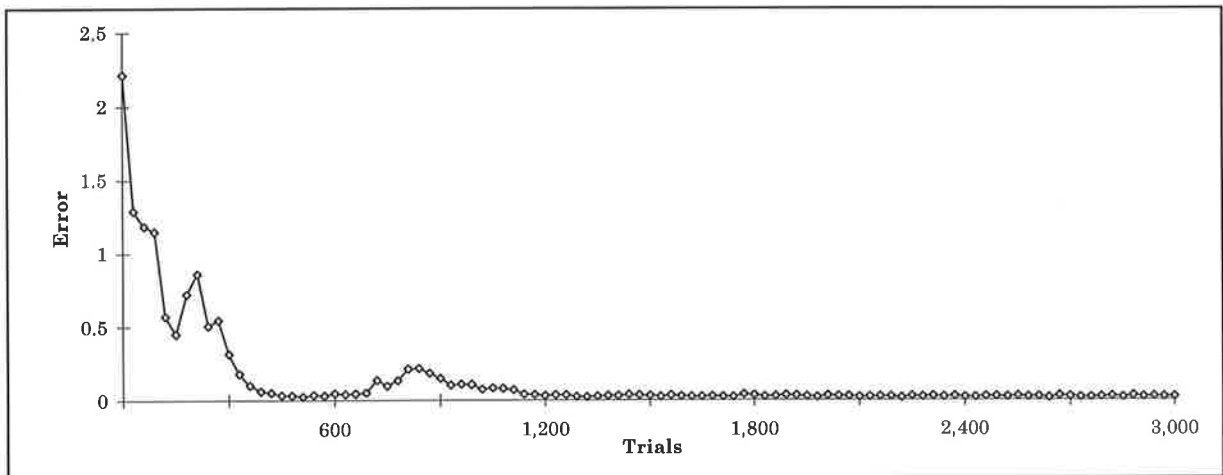


Figure 8.2. The pattern of change of the external error measure across 3,000 trials for the bug model.

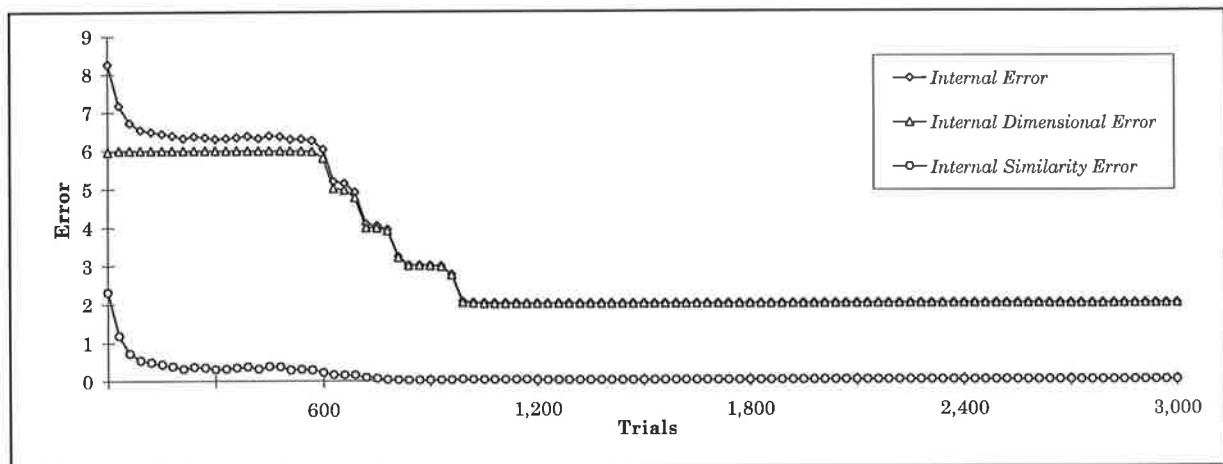


Figure 8.3. The pattern of change of the three internal error measures across 3,000 trials for the bug model.

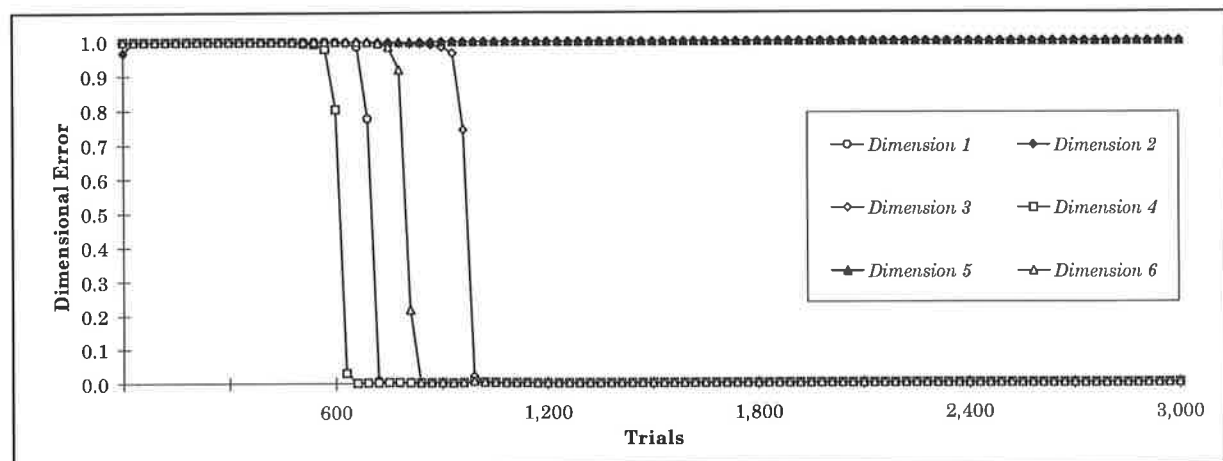


Figure 8.4. The breakdown of dimensional error across the 6 component stimulus dimensions for the bug model.

Once again, the dimensional structure of the internal representation is readily appreciated through an examination of the pattern of change of the dimensional error measure for each possible component dimension. It is evident from Figure 8.4 that the final representational space is two dimensional, consisting of the coordinate values assumed by the six bug types at the second and fifth units in the internal representation layer.

This internal representation is depicted in Figure 8.5, and clearly reflects the sensory properties of the bugs which were deemed to be of importance in simulating the encountered environment. Importantly, the relative spacing of the representations of the different bug types are stable, and accord with the quantitative codings given in Table 8.1. This correspondence demonstrates the model's insensitivity to momentary fluctuations in the sensory properties of the bugs resulting from the addition of noise.

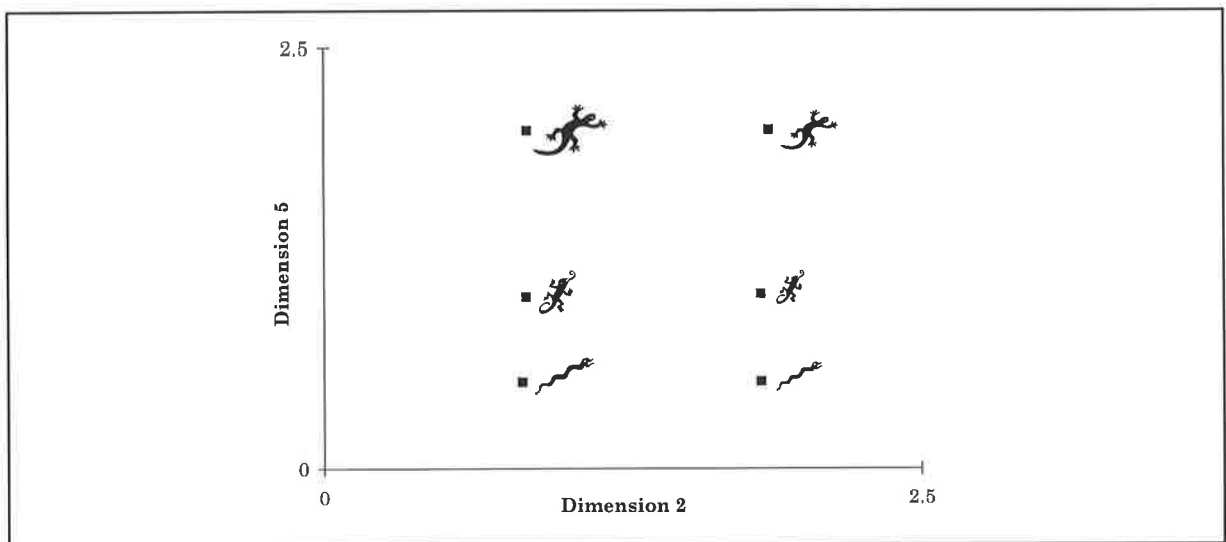





Figure 8.5. The final internal representation learned by the bug model.

### 8.1.2. The Berry Model

A second model in which the model received environmental information relating to sensory properties involved a stimulus set of three berries with various gustatory and olfactory properties, as detailed in Table 8.2. The berry model employed learning rate parameters of  $\lambda_c = 0.01$  and  $\lambda_w = 0.01$ , a  $\beta$  value of 10, and again assumed the City-block metric and altered  $\kappa$  to maintain  $\kappa v$  at 1.5.

Table 8.2. Sensory Properties of Three Berries

			
Gustatory	+2	+1	0
Olfactory	+2	+1	0

Noise was independently added to both of the environmental feedback units, and was again sampled from a Gaussian distribution with zero mean and a variance of 0.1. The effect of the



incorporation of this noise is evident from the sample of sensory feedback for each of the three berry types, as shown in Figure 8.6.

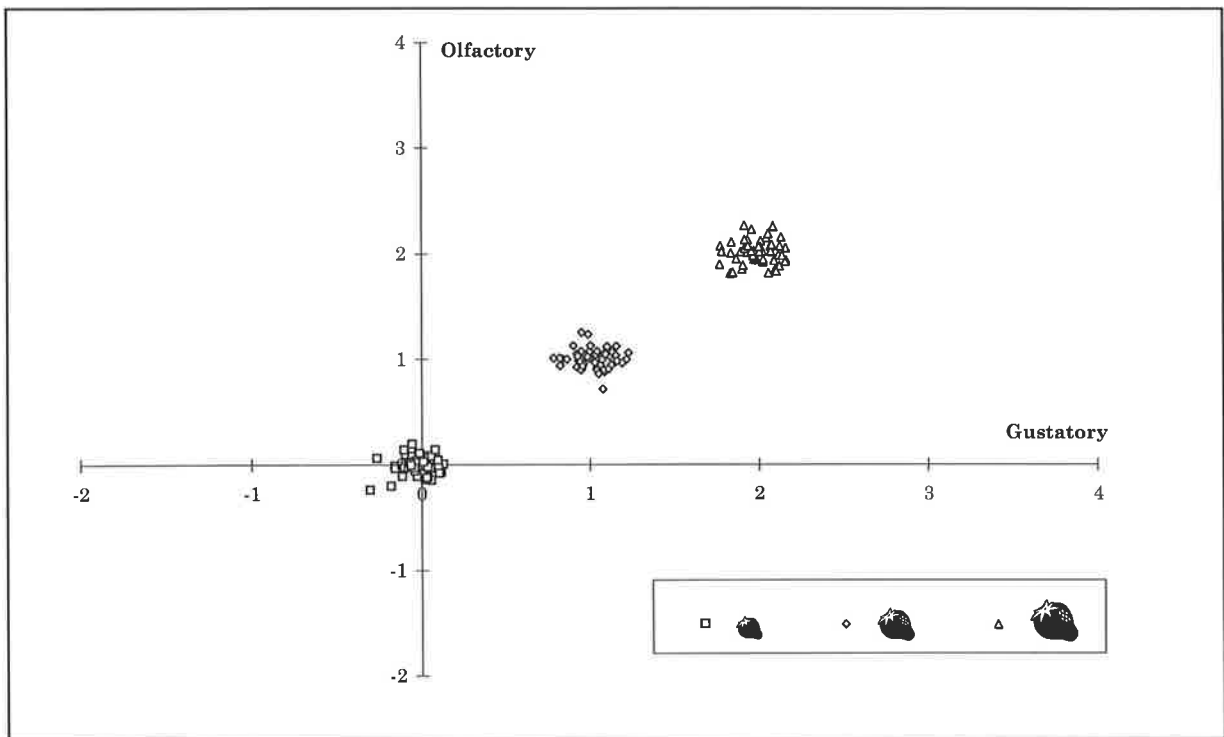


Figure 8.6. The nature of the sensory information received by the berry model. 50 samples for each berry type are shown.

The pattern of change of the external error measure, as shown in Figure 8.7, indicates that within 200 trials the model is able to predict, as accurately as the noise allows, both sensory properties of the three berry types.

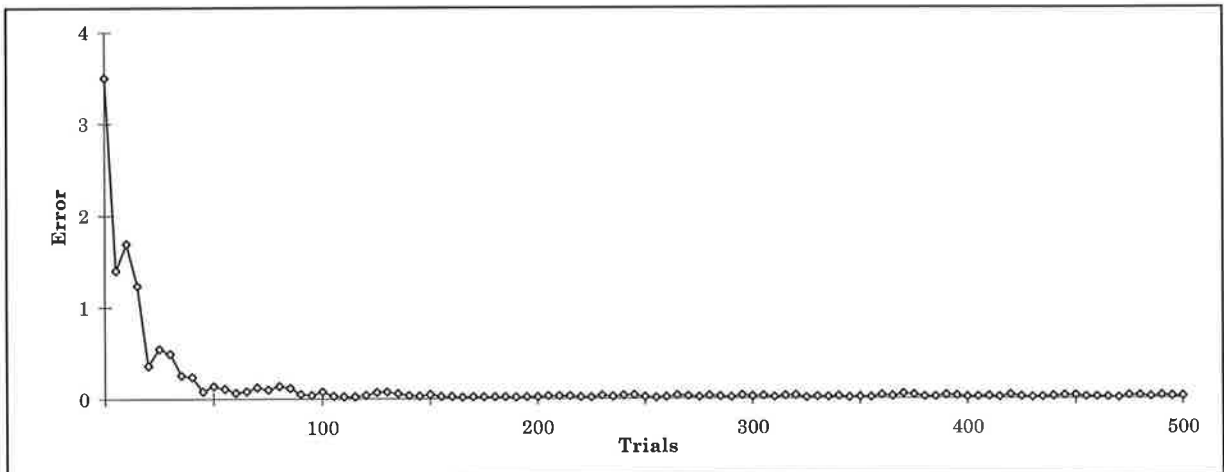


Figure 8.7. The pattern of change of the external error measure across 3,000 trials for the berry model.

Figure 8.8 depicts the various internal error measures, and suggests that after approximately 150 trials the berry model has accommodated the target similarity values, as derived in response space, within a representational psychological space of reduced dimensionality.

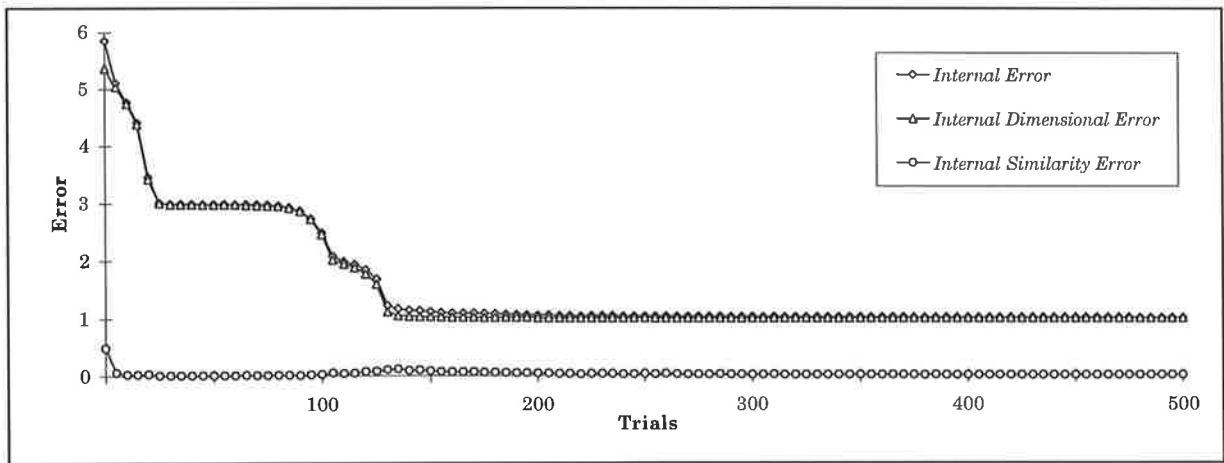


Figure 8.8. The pattern of change of the three internal error measures across 3,000 trials for the berry model.

The breakdown of the dimensional error values of the six component dimensions, shown in Figure 8.9, indicates that the final representational structure is one dimensional, consisting only of the second unit in the internal representation layer.

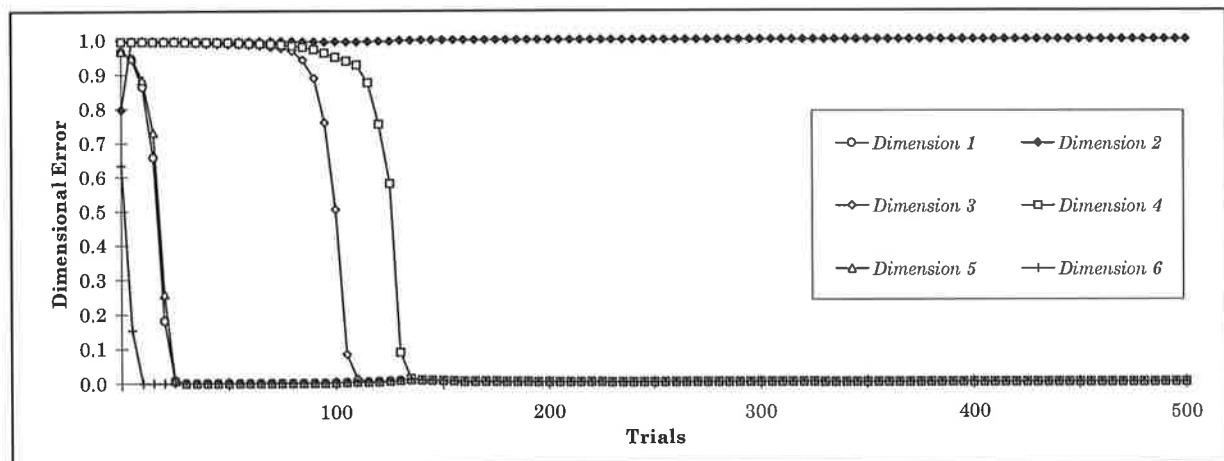


Figure 8.9. The breakdown of dimensional error across the 6 component stimulus dimensions for the berry model.

This representation is displayed in Figure 8.10, and indicates that the model fails to distinguish between the gustatory and olfactory sensory properties of the three berry types. The source of the one-dimensional treatment of the berries relates to the correlation between the sensory properties across the berry types evident in Table 8.2. In the limited environment encountered by the model, the smell of a particular berry type completely determines the taste associated with those berries. Never having experienced a separation of olfactory and gustatory sensation, the model collapses these two properties into a unitary representational whole. Thus, if the gustatory sensation given by the coded property value of -1 is identified with the pain associated in consuming a poisonous berry, then the smell of -1 value for that berry type is directly linked to the pain. In effect, the model considers the smell *is* that of the poison which caused the gustatory pain.

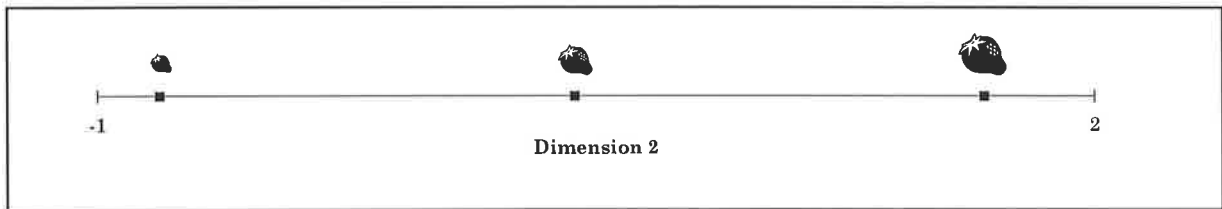


Figure 8.10. The final one dimensional internal representation learned by the berry model..

It is important to emphasise that the integration of sensory properties in this manner is predicated upon complete correlation. If a fourth berry type were to be encountered by the model which had the olfactory sensation of one previous type, and the gustatory effect of another, then the model would develop a two dimensional representation - with axes corresponding to the two sensory properties - across all four berry types. Thus, the model only collapses potentially independent representational dimensions if, within the environment experienced by the model, the maintenance of the dimensional distinction affords no predictive or explanatory advantage.

The model's behaviour, in its elimination of redundant information through the detection of correlation, is reminiscent of Miller's (1962) suggestion that:

“[t]he kaleidoscopic flux of our experience is laced through with correlations we call objects, and it is the task of our perceptual system to discover and identify the correlations, dependencies, and redundancies that signal an object's appearance. In the process of accomplishing this task it seems that the amount of information we are able to handle must be quite small. Because we are limited, the small amount of information that we can handle must be carefully refined to represent just those aspects that are significant for guiding behaviour” (p. 178)

---

## 8.2. Categorical Associations

To examine the model's ability to develop psychological space representations from the environmental constraints implicit in the categorical associations of a stimulus set, three task domains were employed. In each case, the environment was simulated by a binary relation defined between the stimulus set and a set of categorical associations, such that the relation indicated whether or not a particular stimulus was a member of a particular category.

Rather than treat the various categorical associations as stimulus dimensions to be learned as part of a separable psychological space representation, an internal representational structure was sought which viewed the stimuli in terms of the holistic similarity judgments which characterise integral representations. Multidimensional scaling is frequently employed in this way in applications which are not explicitly concerned with the modelling of mental representations (eg. Schiffman, Reynolds & Young 1981). It seems reasonable, however, to apply this approach to the construction of the graded conceptual structures which appropriately model mental representation in many cognitive tasks (eg. Rosch 1978). Indeed, the notion that humans form low-dimensional integral representational structures which summarise the similarity relationships implicit in high-dimensional separable categorical associations has considerable psychological appeal. In particular,

capacity limitations and general requirements of cognitive economy suggest that the development of such representations may constitute the best approach to modelling some stimulus domains.

### 8.2.1. The Bird Model

As a first example of this general approach, the categorical associations of six birds, as given in Table 8.3, were adapted from Smith (1989, Table 13.2).

Table 8.3. Sensory Properties of Six Birds

	flies	sings	lays eggs	is small	nests in trees	eats insects
robin	■	■	■	■	■	■
starling	■	■		■	■	■
vulture	■	■			■	
sandpiper	■	■	■	■		■
flamingo						
penguin			■			

The bird model consisted of six units in the stimulus input, exemplar, and response exemplar layers, corresponding to the six different birds. Six units were also placed in the response and environment feedback layers, in accordance with the various categorical associations. Finally, following the practice adopted in previous models, six units were placed in the internal representation layer. Learning rate parameters of  $\lambda_c = 0.01$  and  $\lambda_w = 0.05$  were used,  $\beta$  was set to 5, and the Euclidean metric was assumed. The information parameter,  $\kappa$ , was adapted to maintain a  $\kappa v$  value 0.5, reflecting the fact that the response space was six-dimensional.

The pattern of change of the external error measure, shown in Figure 8.11, indicates that, after approximately 2,500 trials, the model has learned the categorical associations of all six bird stimuli.

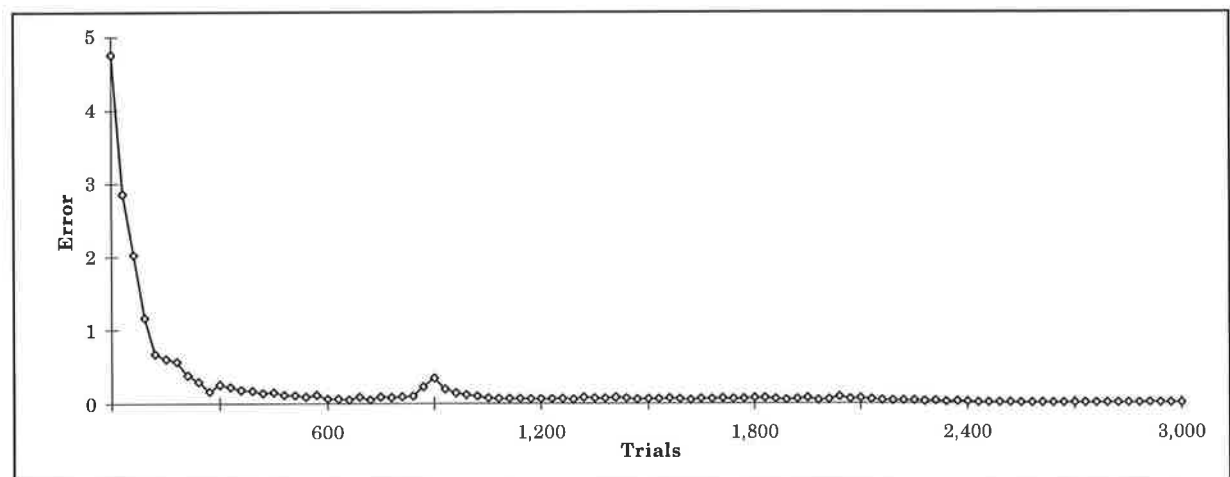


Figure 8.11. The pattern of change of the external error measure across 3,000 trials for the bird model.

The internal error measure, depicted in Figure 8.12, show that the target similarity values arising from these learned categorical associations have been accommodated by the model within a representational space of reduced dimensionality.

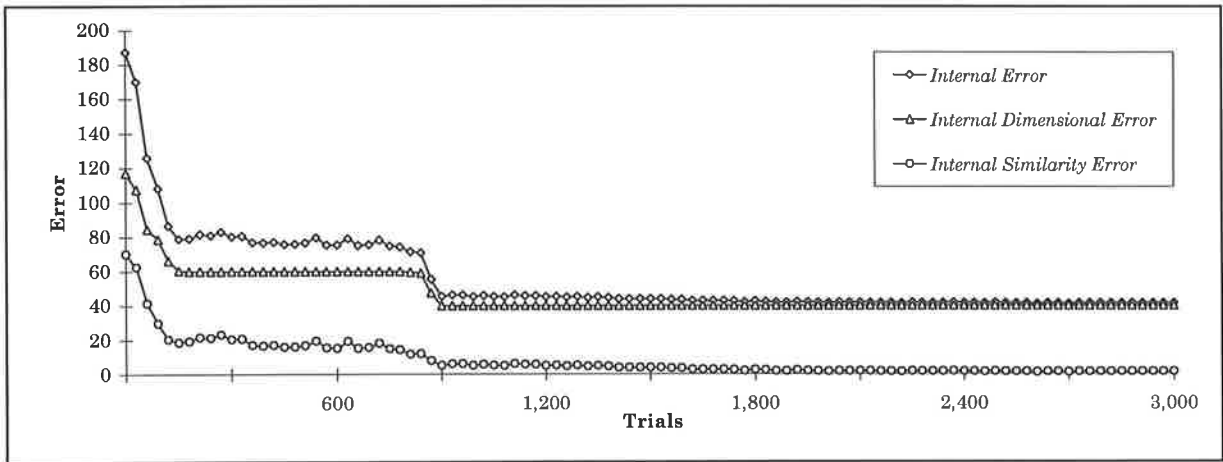


Figure 8.12. The pattern of change of the three internal error measures across 3,000 trials for the bird model. The internal similarity error has been enlarged by a factor of 100 to assist in both its interpretation, and that of the internal error measure.

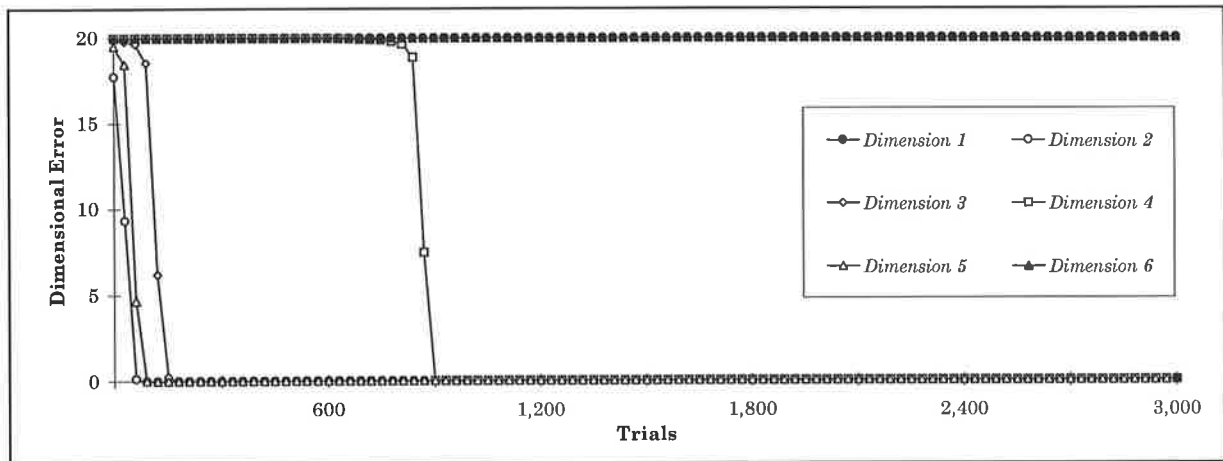


Figure 8.13. The breakdown of dimensional error across the 6 component stimulus dimensions for the bird model.

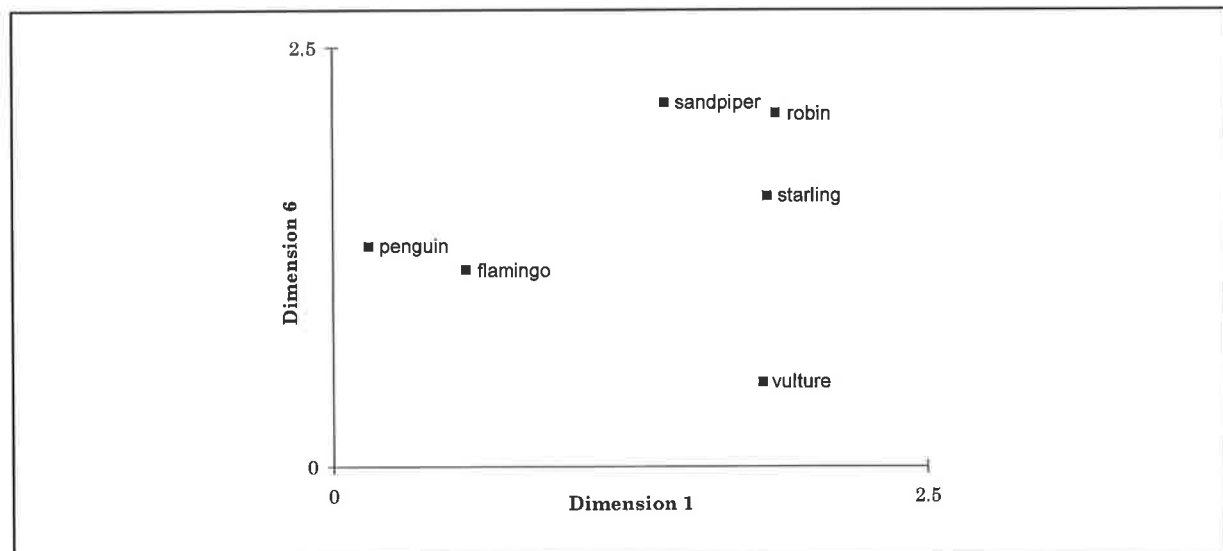


Figure 8.14. The final two dimensional internal representation learned by the bird model.

An examination of the individual dimensions of the internal representation layer, given in Figure 8.13, reveals that the final representation is two-dimensional, consisting of the first and sixth units in this layer. This representational structure is shown in Figure 8.14, and provides evidence of

the appropriateness of the internally derived indices of psychological similarity. In particular, the distance between the various stimulus points seems to reflect an overall similarity between the birds they represent. For example, the fact that ‘vulture’ is relatively separated from the remainder of the stimulus set could be suggested to reflect the relatively large psychological difference between vultures and the other birds. Similarly, the clustering of ‘sandpiper’, ‘robin’ and ‘starling’, and of ‘penguin’ and ‘flamingo’, also accord well with intuition.

### 8.2.2. The Animal Model

A more detailed model of the learning of mental representations through categorical associations extended the bird task domain to incorporate a set of 25 animals. The 14 categorical associations of these animals, as given in Table 8.4, were adapted from the ‘Zoo’ data base (Merz & Murphy 1996).

Table 8.4. Categorical Associations of 25 Animals

	h a i r	f e a t h e r s	e g g s	m i l k	a i r b o r n e	a q u a t i c	p r e d a t o r	t o o t h e d	b a c k b o n e	b r e a t h e s	v e n o m o u s	f i n s	t a i l	d o m e s t i c
antelope	■			■				■	■	■			■	
bear	■			■			■	■	■	■				
chicken		■	■		■				■	■			■	■
crow		■	■		■		■		■	■			■	
deer	■			■				■	■	■			■	
dolphin				■		■	■	■	■	■		■	■	
duck		■	■		■	■			■	■			■	
elephant	■			■				■	■	■			■	
flamingo		■	■		■				■	■			■	
frog			■			■	■	■	■	■				
giraffe	■			■				■	■	■			■	
goat	■			■				■	■	■			■	■
gorilla	■			■				■	■	■				
hawk		■	■		■		■		■	■			■	
lion	■			■			■	■	■	■			■	
lobster			■			■	■							
ostrich		■	■						■	■			■	
penguin		■	■			■	■		■	■			■	
piranha			■			■	■	■	■			■	■	
platypus	■		■	■		■	■		■	■			■	
scorpion							■	■	■	■	■		■	
seal	■			■		■	■	■	■	■		■	■	
sparrow		■	■		■				■	■			■	
stingray			■			■	■	■	■		■	■	■	
swan		■	■		■	■			■	■			■	

The animal model contained of 25 units in the stimulus input, exemplar, and response exemplar layers, 14 units in the response and environmental feedback layers, and six units in the internal representation layer. Learning rate parameters of  $\lambda_c = 0.05$  and  $\lambda_w = 0.1$  were employed,  $\beta$  was set to 4, and the Euclidean metric was assumed. The information parameter,  $\kappa$ , was continually modified to maintain a  $\kappa v$  value 0.3, in accordance with the high dimensionality of the

response space.

As is evident from Figure 8.15, the model required more than 10,000 trials before the external error measure consistently indicated that the model was able to make reasonable predictions regarding the categorical associations of the animals.

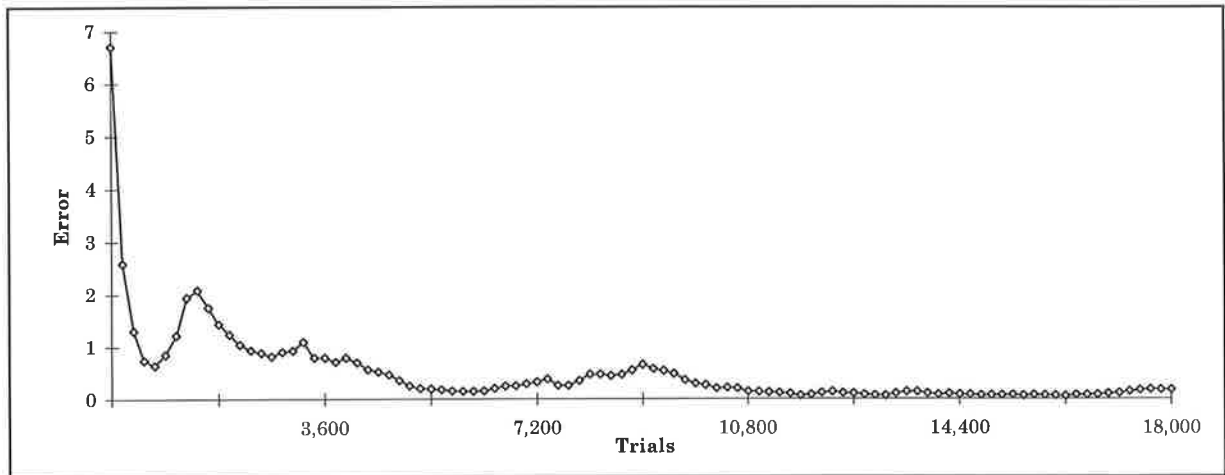


Figure 8.15. The pattern of change of the external error measure across 18,000 trials for the animal model.

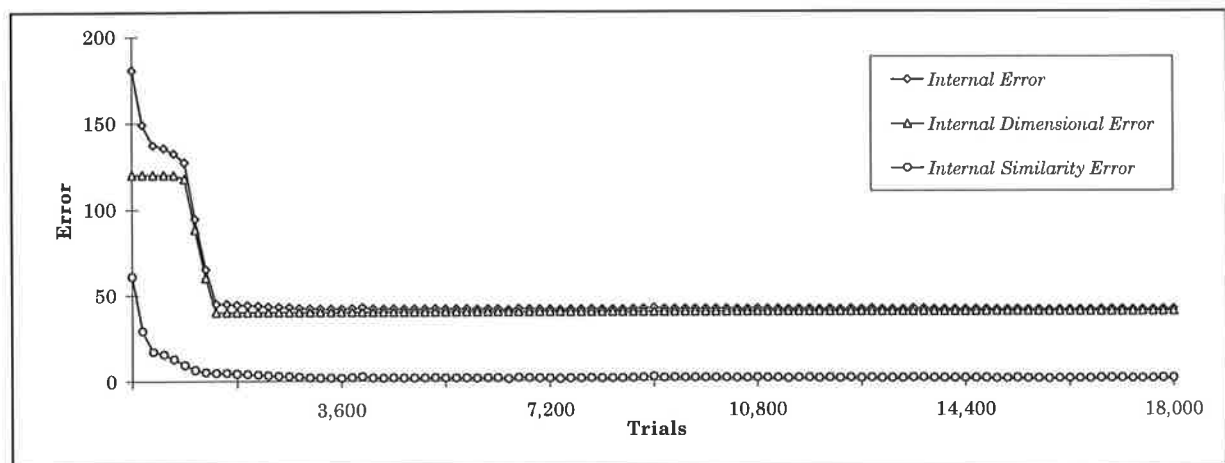


Figure 8.16. The pattern of change of the three internal error measures across 18,000 trials for the animal model. The internal similarity error has been enlarged by a factor of 10 to assist in both its interpretation and that of the internal error measure.

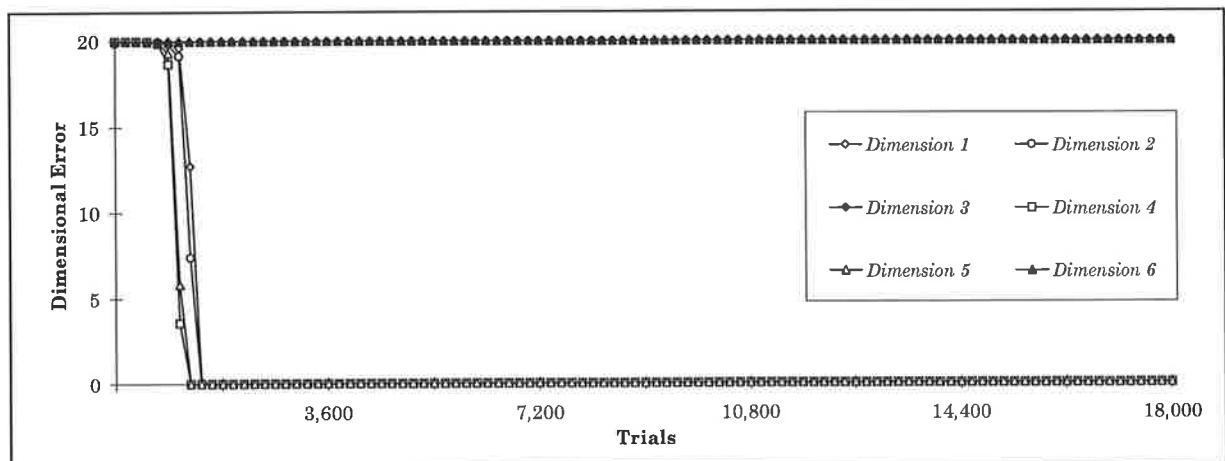


Figure 8.17. The breakdown of dimensional error across the 6 component stimulus dimensions for the animal model.

Throughout this extended learning process, however, the internal error measures shown in

Figures 8.16 and 8.17 suggest that, after the elimination of four representational dimensions, the model was able to accommodate the current target similarity values. Nevertheless, an appropriate psychological space representation cannot be developed until the target similarity values derived in the response space are based upon accurate knowledge regarding the categorical associations of the stimulus set.

This final representational structure is shown in Figure 8.18, and constitutes a more detailed example of the representational appropriateness evident in the bird model. Once again, the proximity of stimulus points in the psychological space reflects the subjective similarity of the various animals, and the discernment of clusters of closely related animals is also possible. For example, the close proximity of 'piranha' and 'stingray' seems appropriate given their shared aquatic and dangerous nature. The birds in the stimulus set - 'crow', 'hawk', 'sparrow', 'ostrich', 'duck', 'flamingo' and 'chicken' - are also clustered together. In addition, 'penguin' is located at the periphery of this cluster, thus creating an appropriately graded structure for the concept 'bird'.

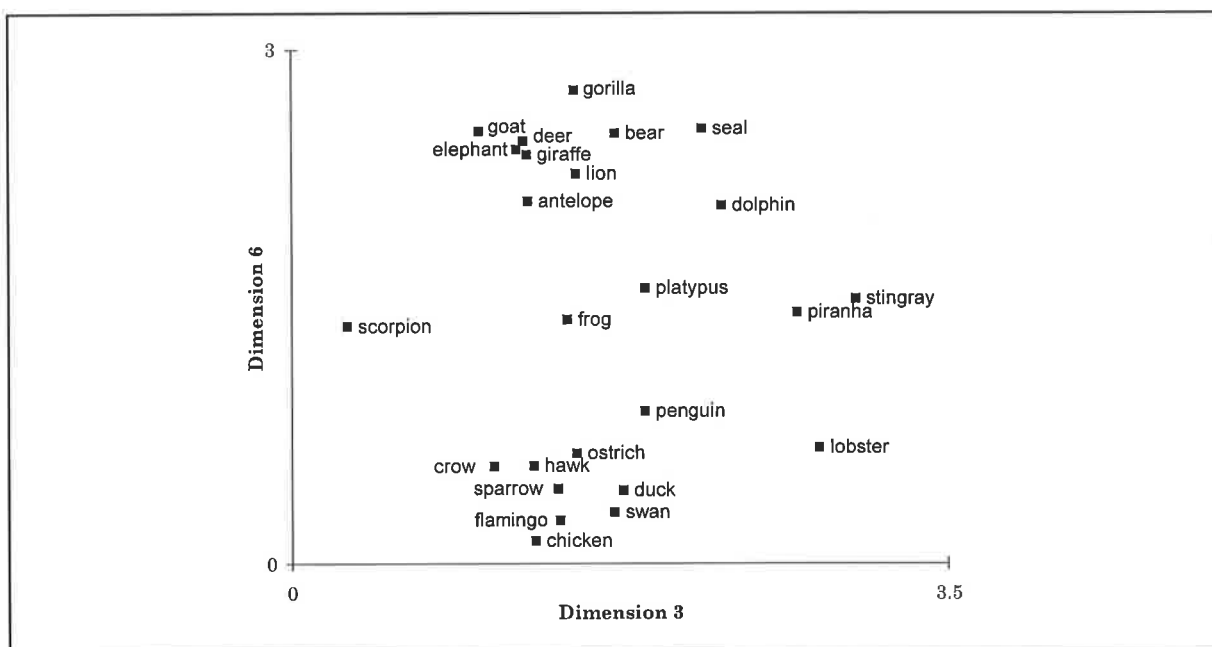


Figure 8.18. The final two dimensional internal representation learned by the animal model.

An evaluation of the representations learned by the animal and bird models can only be made in terms of their intuitive reasonableness. This is unfortunate, since the inability to evaluate the derived representational structure in a more precise manner results in greater justificatory burden being placed on the principles which underpin the model's construction. In particular, establishing the appropriateness of the representations learned by the bird and animal models requires an acceptance of the general psychological space construct, and of the notion that measures of psychological similarity can be generated from the categorical associations of a set of stimuli. Chapters 3 and 6, respectively, constitute detailed attempts to promote such acceptance, and strengthen the inferential chain on which any evaluation of the performance of the current model must rest. Thus, to the extent that the current model is capable of empirical and theoretical



examination, the relevant empirical evidence and theoretical development is presented in those chapters.

In defence of this state of affairs, it is worth recalling that previously developed connectionist models which learn ‘psychological’ internal representations - such as connectionist semantic networks and semantic maps - have been subjected to equally indirect forms of evaluation. As noted in Section 2.3, the semantic map’s derived structure is accepted on the basis of intuitive reasonableness, whereas the representations learned by connectionist semantic networks are typically evaluated through the post-hoc application of clustering or other statistical analysis. There is nothing to be gained by conducting similar analyses upon the representations learned by the current model since, by their method of construction, they adhere to the representational constraints of psychological spaces. As has already been noted in Section 7.4.1, the fundamental distinction between connectionist semantic networks and the current model is that the current model operates in explicit accordance with a psychologically principled theory of human mental representation. The merits of the internal representations developed by the current model reside in their adherence to the empirically and theoretically justified psychological space construct, and a direct evaluation of particular representational structures is, in general, not possible.

### 8.2.3. The Senator Model

A direct examination of the current model’s performance is, however, possible in the situation where a particular representational outcome is known to be appropriate given the environmental feedback the model receives. Such an environment is developed by MacRae (1968, cited in Borg & Lingoies 1987, pp. 172-175) through the specification of the voting patterns of 30 (fictitious) senators across 26 (fictitious) bills, as shown in Table 8.5.

As observed by Borg and Lingoies (1987), these voting patterns imply certain patterns of organisation across groups of senators. In particular, “senators 1 through 5 are ordered in the sense of the simplex ... senator 1 is more similar in his [or her] behavior to senator 2 than to senator 3 ... senators 6,...,10 and 11,...,15 etc. are ordered in the same way” (p. 172). Further exploration reveals that “the various simplexes [sic] discussed above are interrelated in the form of a very regular network” (p. 174), a suggestion which is pursued to analytic confirmation. Specifically, the appropriate geometric representation of the senators consists of a 6 x 5 rectangular lattice with senators 1-5, 6-10, and so on forming the rows/columns, and senators 1,6,11,16,21,26 comprising the first column/row (see Borg & Lingoies 1987, p. 174, Figure 11.1).

To test the ability of the current model to recover this desired representational structure, a senator model was developed, consisting of thirty units in the stimulus input, exemplar, and response exemplar layers, twenty-six units in the response and environment feedback layers, and six units in the internal representation layer. Learning rate parameters of  $\lambda_c = 0.05$  and  $\lambda_w = 0.05$

were used,  $\beta$  was set to 5, the Euclidean metric was assumed, and the information parameter,  $\kappa$ , was adapted to maintain a  $\kappa v$  value 0.3, reflecting the high dimensionality of the response space.

Table 8.5. Voting patterns of 30 senators across 26 bills

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	■	■	■	■			■	■	■				■	■	■					■	■						
2		■	■	■				■	■					■	■						■						
3			■	■					■						■						■						
4				■																							
5																											
6	■	■	■	■		■	■	■	■		■	■	■	■	■				■	■	■						■
7		■	■	■			■	■	■				■	■	■						■	■					■
8			■	■				■	■					■	■						■	■					■
9				■					■						■							■					■
10																											■
11	■	■	■	■		■	■	■			■	■	■	■	■			■	■	■	■					■	■
12		■	■	■			■	■				■	■	■	■					■	■	■					■
13			■	■				■					■	■	■						■	■	■				■
14				■											■	■					■	■					■
15															■						■	■					■
16	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■			■	■	■
17		■	■	■		■	■	■			■	■	■	■	■				■	■	■	■				■	■
18			■	■			■	■				■	■	■	■				■	■	■	■				■	■
19				■				■					■	■	■					■	■	■				■	■
20								■					■	■	■					■	■	■				■	■
21	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
22		■	■	■		■	■	■	■	■	■	■	■	■	■			■	■	■	■	■			■	■	■
23			■	■			■	■			■	■	■	■	■			■	■	■	■	■			■	■	■
24				■				■				■	■	■	■				■	■	■	■			■	■	■
25								■					■	■	■				■	■	■	■			■	■	■
26	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
27		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
28			■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
29				■			■	■			■	■	■	■	■			■	■	■	■	■	■	■	■	■	■
30								■				■	■	■	■			■	■	■	■	■	■	■	■	■	■

The change in the external error measure, shown in Figure 8.19, indicates that, even after 50,000 trials, the model has not learned to predict the voting patterns of the senators, in the sense of producing the precise activation values supplied by the environmental feedback layer. The errors which remain, however, are all caused by activation values being produced which are of the appropriate sign, but of insufficient magnitude to be measured as error-free. Thus, in the sense that any reasonable decision rule which acted upon the response layer would produce voting patterns in accordance with Table 8.5, the model can be regarded to have acquired this information.

An examination of the internal error measure, depicted in Figure 8.20, and its breakdown by stimulus dimension, shown in Figure 8.21, reveal that by the time 20,000 trials have been completed, the target similarity values have been accommodated by the model within a two-dimensional representational space

This final representational structure is shown in Figure 8.22, and may be evaluated in terms of its compliance with the rectangular lattice structure known to be appropriate. Topologically, the model's representation is largely consistent with the desired representational outcome. The location of the stimulus points corresponding to the 30 senators generally accords, in an ordinal sense, with the layout described earlier. For example, the connection of points 6 through 10 does

not intersect with the connection of points 1 through 5, nor with the connection of points 11 through 15, and lies between these two connections. Similar statements can be made regarding much of the remainder of the representation, including connections made along the other axis of the lattice involving, for example, the points 5, 10, 15, 20 and 25.

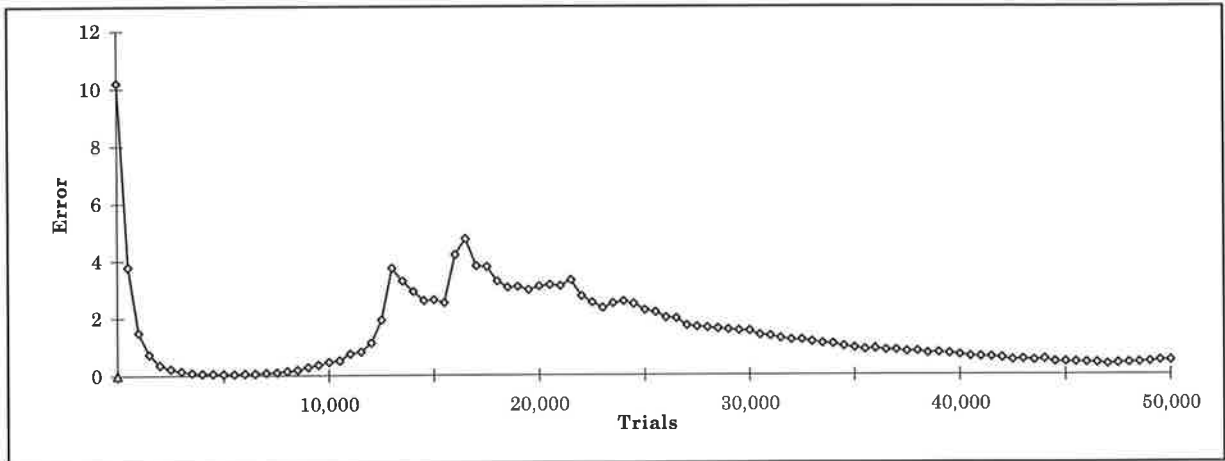


Figure 8.19. The pattern of change of the external error measure across 50,000 trials for the senator model.

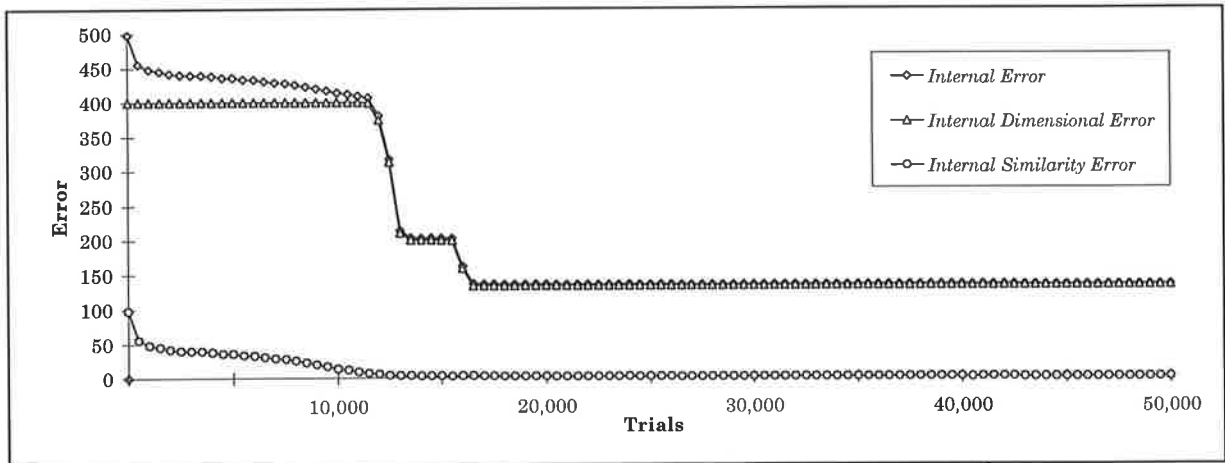


Figure 8.20. The pattern of change of the three internal error measures across 50,000 trials for the senator model.

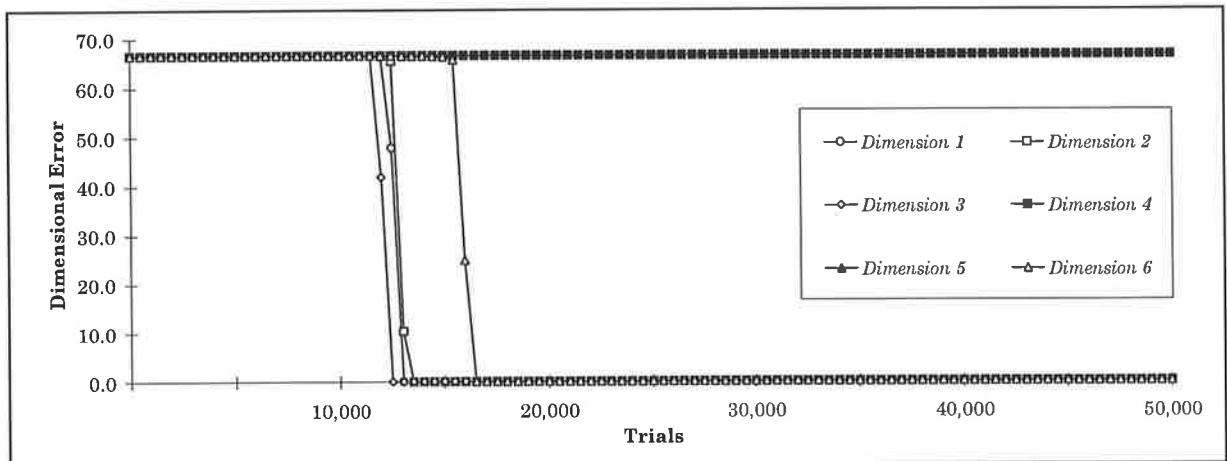


Figure 8.21. The breakdown of dimensional error across the 6 component stimulus dimensions for the senator model.

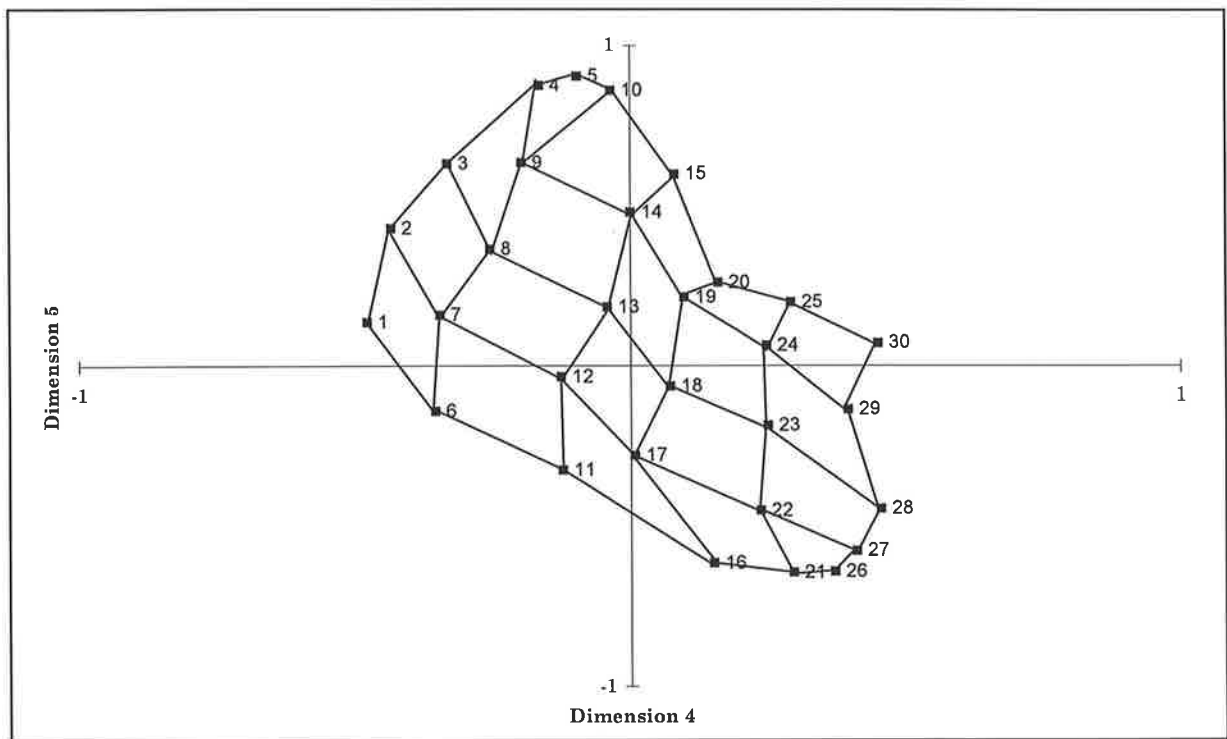


Figure 8.22. The final two dimensional internal representation learned by the senator model.

Clearly, however, some metric information implicit in the environmental feedback received by the model is not accommodated by the learned internal representation. With reference to Figure 8.19, the cause of this deficiency resides in the model failing to generate response output values sufficiently close to +1 and -1. The result of this shortcoming is that, whilst the model remains capable of generating appropriate (binary) predictions regarding the senators' voting patterns, the equivalent importance which must be ascribed to each vote in order to generate a regular lattice is not recognised. In effect, if the vote of a particular senator in favour of a particular bill is learned to be made in the affirmative, but the response output activation which signals this knowledge is significantly less than +1, this vote will exert less influence than it should upon the derived metric representational structure.

Admittedly, there is probably some scope for debate regarding whether or not such metric equivalence is implied by the nature of Table 8.5, which essentially exists at an ordinal level. As emphasised in Chapter 3, however, one of the fundamental tenets of the psychological space position is that mental representational structures are appropriately modelled by attempting to recover metric structure from ordinal information. Thus, the failure of the senator model to derive an entirely regular lattice structure should be interpreted as a weakness. Given the topological success of the model, however, this weakness is relatively minor and is not indicative of fundamental shortcomings of the model's operation. Indeed, it seems reasonable to suggest that a reduction in learning rate parameters, possibly coupled with an increased number of learning trials, would be likely to result in the required metric information being successfully internalised. Consequently, the external error would become negligibly small, and the regular rectangular lattice

representational configuration would be recovered.

### 8.3. Adaptation To A Dynamic Environment

Section 6.2.1 identified the practice of developing static models of the environment as a primary weakness of attempts to impose external representational constraints upon connectionist networks. All of the preceding demonstrations of the current model have, however, been based on essentially static characterisations of the environment. Whilst Section 8.1 incorporated stochastically varying feedback, this variation was super-imposed upon the world models, presented in Tables 8.1 and 8.2, which remained fixed during the model's operation.

#### 8.3.1. The Dynamic Bug Model

To examine the mental representation learning model's ability to derive appropriate internal representations in a dynamic environment, the thermal and auditory sensory properties of six bugs were defined to assume the values given in Table 8.6 for 3,000 encounters, and then to adopt the values given in Table 8.7. This scenario may be conceived as one corresponding to a model of a world in which, after a certain time, the bugs which previously had varied with regard to the noises they made, become mute.

Table 8.6. Initial Sensory Properties of Six Bugs













						
Thermal	0	-1	+1	0	-1	+1
Auditory	+1	+1	+1	0	0	0

Table 8.7. Final Sensory Properties of Six Bugs

						
Thermal	0	-1	+1	0	-1	+1
Auditory	0	0	0	0	0	0

The dynamic bug model was of identical construction to the bug model described in Section 8.1.1, except that the value of  $\beta$  was set to 15, and was operated for 6,000 trials.

The pattern of change of the external error measure across these trials is shown in Figure 8.23, and demonstrates that the initial sensory properties of the various bugs are learned within about 1,500 trials. Of particular interest is the rise in the external error immediately after 3,000 trials have been completed, reflecting the incompatibility of the model's knowledge at that time with the information it is receiving from environmental feedback. Clearly, however, the model is able to learn to predict correctly the new sensory properties of the bugs, since the external error quickly achieves a negligible value which it maintains.

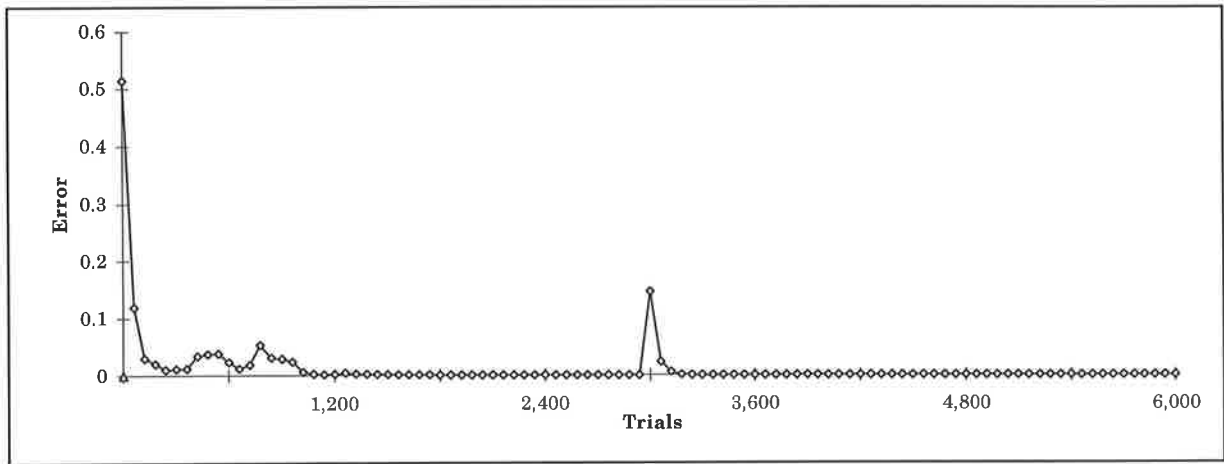


Figure 8.23. The pattern of change of the external error measure across 6,000 trials for the dynamic bug model.

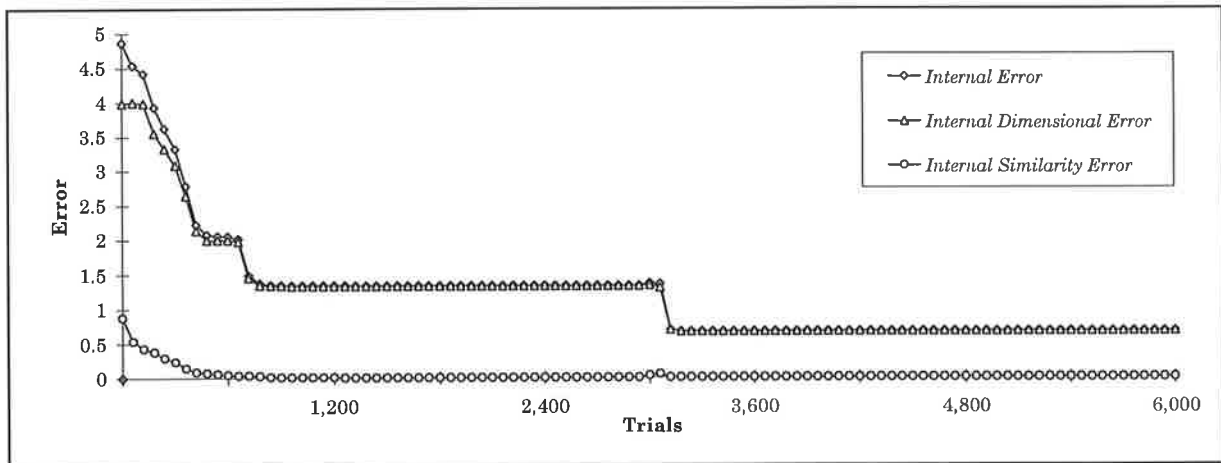


Figure 8.24. The pattern of change of the three internal error measures across 6,000 trials for the dynamic bug model.

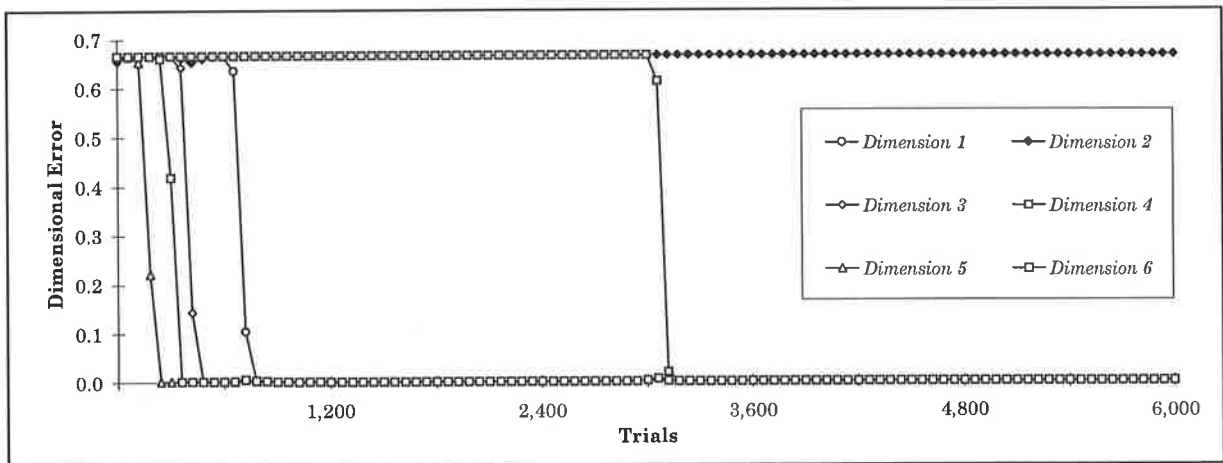


Figure 8.25. The breakdown of dimensional error across the 6 component stimulus dimensions for the dynamic bug model.

The pattern of change of the internal similarity error measure, shown in Figure 8.24, suggests that the model accommodates the internalised psychological similarities after 1,000 trials and is further able to adapt its representational structure to accommodate the new environment it encounters. The internal dimensional error, also shown in Figure 8.24, indicates that the initial representational space is two dimensional whilst, after the environment alters at trial 3,000, a stimulus dimension is removed, resulting, ultimately, in a one dimensional representational

structure.

The breakdown of the internal dimensional error measure in terms of the individual units in the internal representation layer, given in Figure 8.25, reveals that the two dimensional configuration consists of stimulus dimensions two and four, with dimension four being removed following the adaptation of the model's knowledge to the altered environmental information.

The stable internal representation derived by the dynamic bug model between trials 1,500 and 3,000 is shown in Figure 8.26. With reference to Table 8.6, it can be seen that, as with the demonstrations given in Section 8.1, this representational structure appropriately reflects the sensory information available to the model at this time.

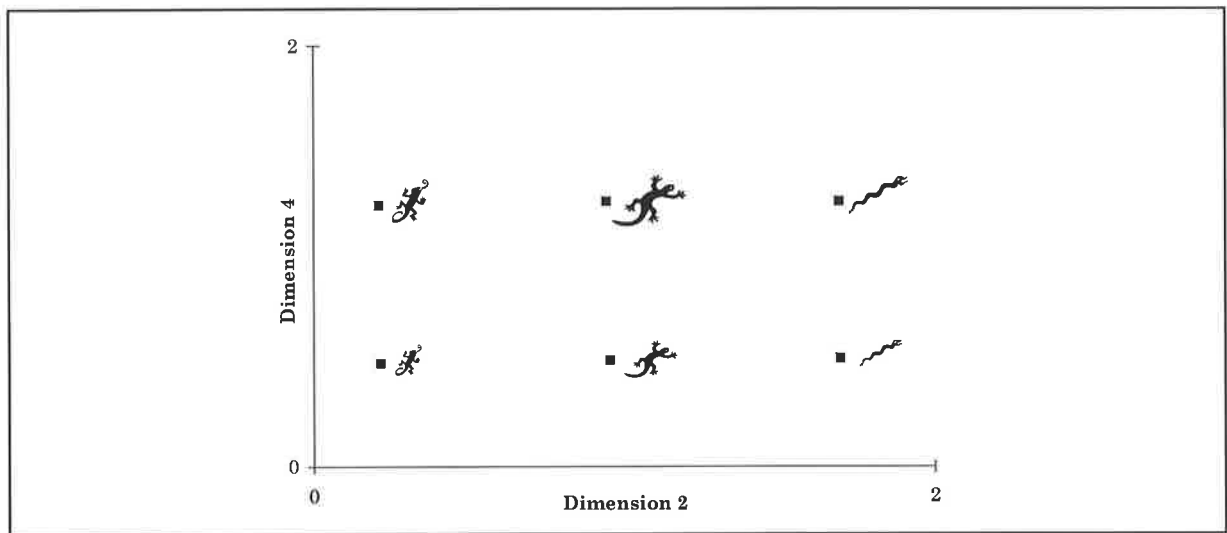


Figure 8.26. The two dimensional internal representation of the dynamic bug model between 1,500 and 3,000 trials, before the environmental feedback is altered.

Similarly, the stable representational configuration derived after the sensory properties of the bugs has been altered, which is displayed in Figure 8.27, recognises the lack of information provided by considering the auditory properties of the bugs, and constitutes a one dimensional representation corresponding solely to their thermal properties. Effectively, since all of the bugs in the new environment are mute, the model discards its auditory receptors as a source of information capable of usefully distinguishing between the bugs it encounters.

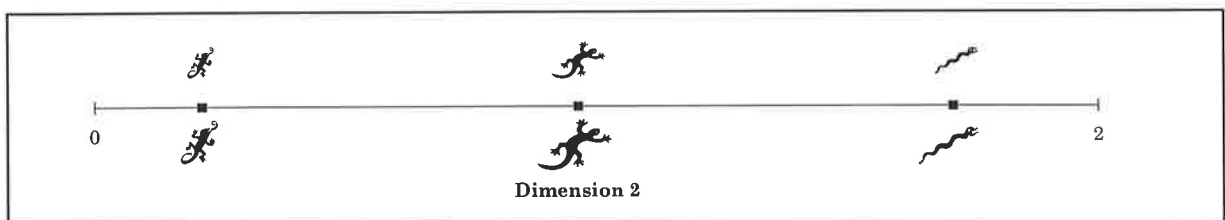


Figure 8.27. The final one dimensional internal representation learned by the dynamic bug model.

### 8.3.2. Adaptational Possibilities And Interpretations

There is no reason why the model's adaptation to a dynamically changing environment must involve the removal of stimulus dimensions. Some environmental changes may be able to be absorbed within a psychological space of the dimensionality already established, whilst others may

require the incorporation of additional stimulus dimensions. In principle, the mental representation learning model can accommodate both of these scenarios. Realising a different representational structure across the same number of internal representation units is straightforward, whilst the fact that units in this layer are never permanently removed means that they can assume a representational role at any stage during the model's operation. In terms of the conjugate space analysis (recall Figure 4.4), stimulus dimensions which lie on the line of non-contribution may be moved from this line if the demands of the similarity error outweigh those of the dimensional error. The primary way in which such demands might arise is through a change in the environmental properties of the stimuli, which creates a different set of similarity relationships for the similarity error to satisfy.

It is worth noting that the ability of the model to incorporate extra dimensions potentially circumvents the arbitrariness of initially placing 6 units in the internal representation layer (recall Section 4.3). Since the model can construct or remove internal representation units as required, there is little reason to attempt to determine a principled upper bound upon the number of dimensions required for the psychological space representation of a stimulus set. Conceptually, the model may be considered as operating with an infinite upper bound on the number of internal representation units. The only alteration that is required, therefore, is the installation of a mechanism which creates as many internal representation units beyond six as are demanded to minimise the total error measure. Under this slightly extended scheme the model is simply enacting 'pruning' and 'growing' techniques of the form previously employed in connectionist modelling (see Ash & Cottrell 1995, Reed & Marks 1995 for summaries).

It is interesting to note that, when the model does alter its internal representational structure through the addition and removal of stimulus dimensions, the adaptive process is somewhat reminiscent of the stage-wise developmental theory espoused by Piaget (1970, see also Lewis 1994, van der Maas & Molenaar 1992). Under this view, human cognitive development involves a series of abrupt changes punctuating periods of stability, rather than a continual process of gradual improvement. These sudden changes arise from the learning process of equilibration which seeks to balance assimilatory and accommodatory tendencies (recall Section 7.4.2). The equilibrium usually achieved by this process is disturbed when the assimilation of environmental information is impossible, even with the action of accommodatory processes. In such circumstances, a dramatic restructuring of underlying mental representational structure is required, and it is the emergence of this new structure which corresponds to cognitive development.

The type of fundamental change involved in stage-wise development could, at least metaphorically, be likened to modification of the dimensionality of a psychological space representation. Whilst, as detailed in Section 7.4.2, assimilation corresponds to updating the



model's learned associations, and accommodation corresponds to the adjustment of the internal representations of stimuli, the modification of the nature of the psychological space underpinning these representations constitutes a more significant event. Changing the dimensionality of a previously stable psychological space heralds the adoption of a fundamentally different representational approach, since, unlike moving the location of a single stimulus, it immediately affects the entire set of similarity relationships embodied within the space.

The difference in the environmental properties of the stimuli encountered by the dynamic bug model creates inconsistencies in the model's representation which are best removed through a dimensional change. More specifically, the absence of noise emanating from the bugs after 3,000 trials is incompatible with the auditory dimension of the psychological space representation which has been learned. In a Piagetian developmental context, therefore, the rapid removal of the fourth stimulus dimension loosely corresponds to a fundamental reorganisation of the model's internal representational structure to resolve tensions or stresses between this structure and incoming environmental information.

## Chapter 9: Extensions To The Mental Representation Learning Model

This chapter considers various extensions to the model of mental representation learning described in Chapter 7 and evaluated in Chapter 8. First, a series of refinements to the model are discussed. Secondly, preliminary investigations are presented of a more significant extension which allows the model to encounter continuous stimulus domains through learning psychophysical mappings.

---

### 9.1. Model Refinements

The various refinements to the model described in this section constitute modifications the architectural, processing and learning foundations developed in Chapter 7. Most of these refinements have been alluded to in previous chapters, and many have already been preliminarily canvassed. Having now completed the demonstration and evaluation of the final model, however, the reconsideration and consolidation of these suggested improvements appears worthwhile.

#### 9.1.1. Adaptive Parameter Setting

All of the parameters incorporated within the model, with the exception of the information parameter  $\kappa$ , assume fixed, pre-determined values throughout the model's operation. The information parameter, in contrast, is continually adaptively modified on the basis of entropic considerations detailed in Section 7.3, since its value is capable of significantly influencing the model's behaviour. Whilst, for the other parameters, some effort has been made to demonstrate that the results attained by the model are relatively insensitive to their precise values, there remains considerable scope for improving the model by allowing these values to change in systematic ways during the course of the model's operation.

The possibility of adaptively modifying the dimensionality reduction parameter of the connectionist multidimensional scaling model was discussed in Section 5.2.1, where it was suggested that appropriate changes might be derived from the pattern of change of the dimensional error measure. The subsequent extension of the multidimensional scaling model to learn environmentally constrained mental representations provides further impetus for the development of such a technique. Through being capable of continually changing the dimensionality reduction parameter in accord with the internal dimensional error measure, the ability of the model to learn and modify a mental representational structure whilst encountering a dynamically changing environment may well be enhanced.

The learning rate parameters incorporated within the current model, which serve to adjust the model's prediction of the environmental properties of stimuli, and adjust the internal representations of those stimuli, could also be subjected to adaptive modification during the course

of the model's operation. Such modification could potentially enhance the reliability and efficiency with which the learning rules appropriately alter the various connection weights on which they act. Once again, the information necessary to make these changes would appear to be contained within the pattern of change of the error measures which are being minimised by the learning rules associated with these parameters.

In one sense, incorporating adaptive modification of these learning rate parameters simply constitutes an attempt to improve the model as a learning machine. More fundamentally, however, this extension offers the hope that the time course of human learning of mental representations might be effectively modelled. The current model attempts only to develop stable psychological space representational structures on the basis of modelling the relationship between the human mind and the external world. In seeking to develop a more detailed model which accounts for the momentary changes in mental structures following environmental experience, it would seem likely that the flexibility afforded by adaptively changing learning rate parameters might be required.

### 9.1.2. Selective Attention Response Space Weighting

Section 7.4.1 discussed the possibility of extending the model by introducing selective attentional mechanisms which act to 'stretch' and 'shrink' the axes of the psychological space representational structure (see Kruschke 1992, Nosofsky 1984). Given the architectural equivalence of the psychological and response spaces, the addition of a similar mechanism to the response space is straightforward, and is consistent with the operation of the City-block distance metric. More importantly, the scaling of axes in the response space would seem to have a natural and potentially useful psychological interpretation. Given that the various dimensions of response space correspond to the environmental information being learned by the model, the association of scaling factors would seem to model the conferral of measures of salience or importance upon the sensory properties and categorical associations involved.

There are at least two situations in which the ability to accommodate the relative importance of environmental information in this way might be desirable. First, in relation to sensory properties, it seems reasonable to suggest that there is some adaptive advantage in innately registering some measure of relative salience between different sensory experiences. It could also be argued, for example, that the primacy of vision as a human modality implies that a greater sensitivity or resolution with regards the visual sensory properties, rather than, say, the olfactory properties, of stimuli is desirable. The inclusion of fixed weightings associated with each axis of response space would accommodate precisely this form of 'hard-wired' distinction.

Secondly, differential weights could be introduced to response space axes which correspond to categorical associations. These weights would influence the summary conceptual structures, such as the birds and animals presented in Chapter 7, formed in the model's psychological space. For

example, the categories 'predator' and 'venomous' in Table 8.4 might be assigned relatively greater emphasis than the other associations. These weightings, in distinction to the sensory domain, are not innate, since the categories involved are learned over the time scale of individual lifetimes, rather than on an evolutionary scale. Thus, any consistency across individuals in the assignment of saliencies should be regarded, once more, as the outcome of the application of general adaptation tendencies to the experiences of those individuals. Furthermore, the possibility remains that different individuals confer different sets of weightings to the various categorical associations. Potentially, the different representations resulting from any such 'subjective' saliencies might be reconciled with individual differences in conceptual structures, of the type explored by INDSCAL (see Shepard 1980) and related multidimensional scaling techniques.

From a modelling perspective, the introduction of response space dimensional weightings has the additional advantage of allowing the code through which environmental information is conveyed to be normalised. Particularly with regard to sensory properties, the ability to quantitatively confine a continuum of stimulation within, say, the unit interval, and then accommodate various inter-sensory differences with response space weightings is highly desirable. This approach to coding environmental feedback would help to reduce the somewhat arbitrary nature of the way in which the environment was modelled in Sections 8.1 and 8.3, and would also introduce explicit measures of dimensional salience which might well have predictive and explanatory significance in terms of understanding the relationship between environmental experience and mental representation.

### 9.1.3. Learning The Metric Structure Of Psychological Space

The models described in Chapters 4 and 7 are both forced to assume the operation of one of the Minkowski family of distance metrics within the psychological space representations they learn. Whilst consequential regions afford considerable insight regarding the way in which the interaction of internal stimulus dimensions corresponds to these various metrics, they do not prescribe an obvious method by which the model might be extended to learn an appropriate metric structure. As was noted in Section 4.1.3, the notion of 'canonical distance metrics' developed by Baxter (1996) is more promising in this regard. Since the inability of the model of mental representation learning to self-determine the distance metric it employs could be regarded as one of its most fundamental deficiencies, it seems worthwhile further exploring the potential application of the canonical distance metric formalism.

Recall that the guiding principle for the construction of a canonical distance metric is that representative points in an 'input' space are not *a priori* subject to a similarity metric. Rather, the similarity between pairs of stimuli - which Baxter (1996) terms 'subjective' similarities, but which could equally appropriately be labelled 'psychological' similarities - are determined by the

environment of functions which are learned across the representational input space. Specifically, stimulus points which are, in some sense, similarly classified by the various functions are regarded as similar and are afforded a relatively smaller inter-point distance by the derived canonical distance metric. In general, the environment functions take the form of classificatory mechanisms which assign stimuli to various categories of importance to the model. Clearly, such functions are closely related to the notion of consequential regions, as both essentially encapsulate regions of the representational space which correspond to important environmental kinds.

More formally, following the notation of Baxter (1996, p. 6), the definition of the canonical distance metric may be characterised as follows. Given a representational input space (ie. a psychological space) denoted  $X$ , a set of functions (ie. consequential regions) to be learned denoted  $F$ , then the classifications of two representational points  $x, x' \in X$  for a function  $f \in F$  are given by  $f(x)$  and  $f(x')$ . Further assume that a function  $\sigma: Y \times Y \rightarrow \mathbb{R}$  determines the difference between these classifications (ie. imposes a metric upon the response space), and that a probability measure  $Q$  determines the probability (ie. some form of salience) of each of the classificatory functions. Then, the canonical distance metric, denoted here by  $\xi$ , defines the distance between  $x$  and  $x'$  to be:

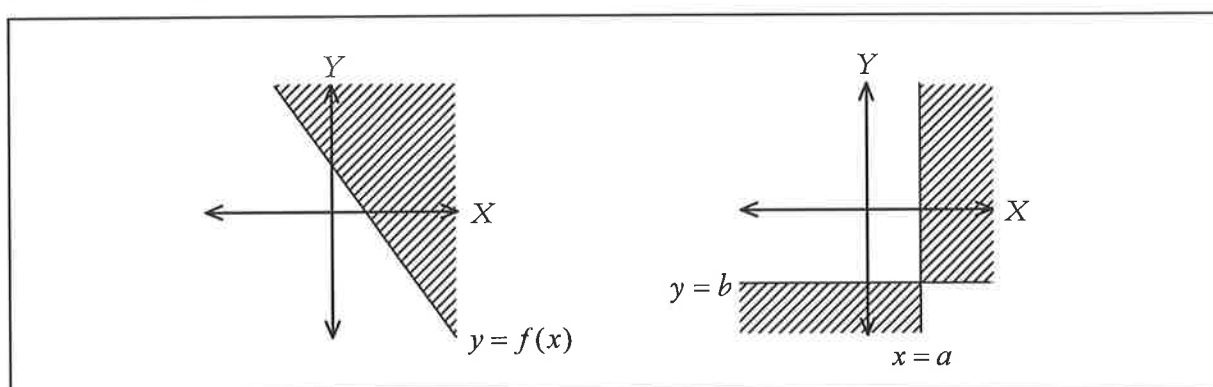
$$\xi(x, x') = \int_F \sigma(f(x), f(x')) \cdot dQ(f) \quad (9.1)$$

Thus, the distance between the two representational points in psychological space is determined by integratively measuring their similarity across the set of environmental properties, weighted according to their salience. Stimuli with common classifications within the environment in which the model operates are placed near each other by the canonical distance metric, whilst stimuli with different classifications are assigned large inter-point distances.

The motivations underpinning canonical distance metric measures and psychological space representations would appear to share much in common. Both attempt to develop representational structures on the basis of similarity measures derived from the adaptively learned environmental properties of those stimuli. Multidimensional scaling focuses upon manipulating the location of stimulus points to achieve such representational structures, whilst canonical distance metric research develops distance metrics which reflect stimulus similarity. The possibility arises, therefore, that the canonical distance metric formalism could profitably be applied to the learning of psychological space distance metrics.

As suggestive preliminary evidence that an ecumenical integration may indeed be achievable, consider the relationship between canonical distance metrics and the separable/integral stimulus distinction. Figure 9.1 depicts two two-dimensional representational spaces containing classificatory functions which identify shaded regions of consequence to the model. The classifier

on the left takes the form of a linear threshold function  $y = f(x)$ , whilst the classifier on the right requires description as a Boolean combination of linear threshold functions in the form  $x > a$  XOR  $y < b$ . The linear threshold function environment is shown by Baxter (1996) to imply a canonical distance metric based on the angle between representative points, a measure which is closely related to the Euclidean distance metric which characterises integral stimuli. The incorporation of Boolean combinations of linear threshold functions, however, seems likely (Baxter, personal communication, October 1996) to correspond to the City-block metric associated with separable stimuli. It certainly appears reasonable to suggest that both of the scenarios depicted in Figure 9.1 might correspond to the learning of functions which confer some form of adaptive advantage. It is, therefore, reassuring that the two distance metrics with significant psychological precedent emerge from these functions.



*Figure 9.1. Contrasting two-dimensional psychological spaces in which decision boundaries are described, on the left, by a linear threshold function, and, on the right, by a Boolean combination of linear threshold functions.*

Whilst this observation is promising, however, there is clearly a need for a far more detailed examination of the relationship between canonical distance metrics and consequential region structures before a connectionist mechanism which learns appropriate distance metrics in developing psychological space internal representations can be formalised.

#### 9.1.4. Response Space Basis Function

The response space basis function given by Equation 7.3 generates target indices of psychological similarity from environmental information internalised in response space. Clearly, therefore, the values produced by this function have the potential to influence the representational structures derived by the model. As discussed in Section 7.1.2, however, the exponential decay form chosen for this function does not have the same strong theoretical and empirical underpinnings of its psychological space counterpart.

One means by which this situation might be redressed is through further empirical study of response generalisation, of the generic type undertaken by Shepard (1958a, Experiment II) and Noble and Bahrick (1956). Given improvements in multidimensional scaling techniques since this research was undertaken, particularly in the form of non-metric algorithms, it is possible that the

form of the response generalisation function might be revealed by such data.

On the theoretical front, there would appear to be considerable merit in examining the sensitivity of the psychological spaces derived by the model to the form of the response space basis function. This could be done by simply employing other functions which meet the requirements of Equations 7.2, and comparing the resultant psychological space representations. Observed insensitivity of the representations to the basis function would remove the imperative to discover *the* generalisation function, whilst differences in the representations could be employed in relation to empirical data to further constrain the set of appropriate functional forms.

A more enterprising theoretical development of the model which addresses shortcomings of the current response space basis function involves the incorporation of notions from non-metric multidimensional scaling. Radial basis functions assume that a metric relationship exists between the argument (distance in response space) and the output (target similarity). Whilst the Universal Law of Generalization justifies this assumption within psychological space, the essence of the difficulties in deriving a basis function is that no such firm guarantee exists in response space. Recalling that the ability of non-metric multidimensional scaling to reveal metric generalisation gradients implicit in ordinal similarity data contributed to the development of the Universal Law of Generalization, it would appear sensible to incorporate this ability within the current model. That is, the location of response units within response space could be used to generate ordinal level data, in the form of topological orderings, rather than distance measures, thus removing entirely the need for a response space basis function.

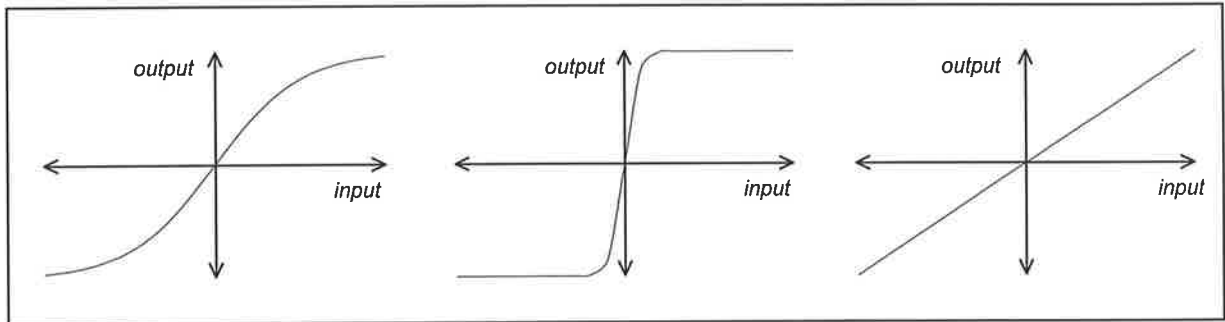
One way in which the current model might be extended in this way is evident in Schneider's (1992) approach to multidimensional scaling, which allows metric and non-metric techniques to be treated as two ends of a continuum. This approach hinges upon the properties of a parameterised sigmoid function, of the form:

$$f_{\alpha}(x) = -\frac{1}{2} + \frac{1}{1 + e^{-\alpha x}} \quad (9.2)$$

where the parameter  $\alpha$  is positive. As is shown in Figure 9.2, different values  $\alpha$  correspond to what might be described as qualitatively different forms of the function  $f_{\alpha}$ . Whilst intermediate values of  $\alpha$  result in the characteristic sigmoid shape, very large values of  $\alpha$  give rise to a curve closely approximating a step function, and very small values of  $\alpha$  correspond to a functional form which is essentially linear. Examples of these three cases are shown on the left, middle and right, respectively, of Figure 9.2.

The insight provided by Schneider's (1992) approach is that, by applying the function  $f_{\alpha}$  to paired similarity data, variations in the parameter  $\alpha$  impose different levels of measurement upon these pairwise discrepancies. In particular, a small value of  $\alpha$  performs only a linear scaling upon

the differences, thus embodying metric assumptions, whilst a large value of  $\alpha$  corresponds to non-metric multidimensional scaling, since only the rank order of the similarity values is captured in the effectively discrete output of the sigmoid. Choosing values of  $\alpha$  between these two extremes, therefore, allows for a continuum of multidimensional scaling techniques encapsulating measurement assumptions which are stronger than ordinal, without being fully metric.



*Figure 9.2. The form of the parameterised sigmoid over intermediate (left), large (middle), and small (right) positive parameter values.*

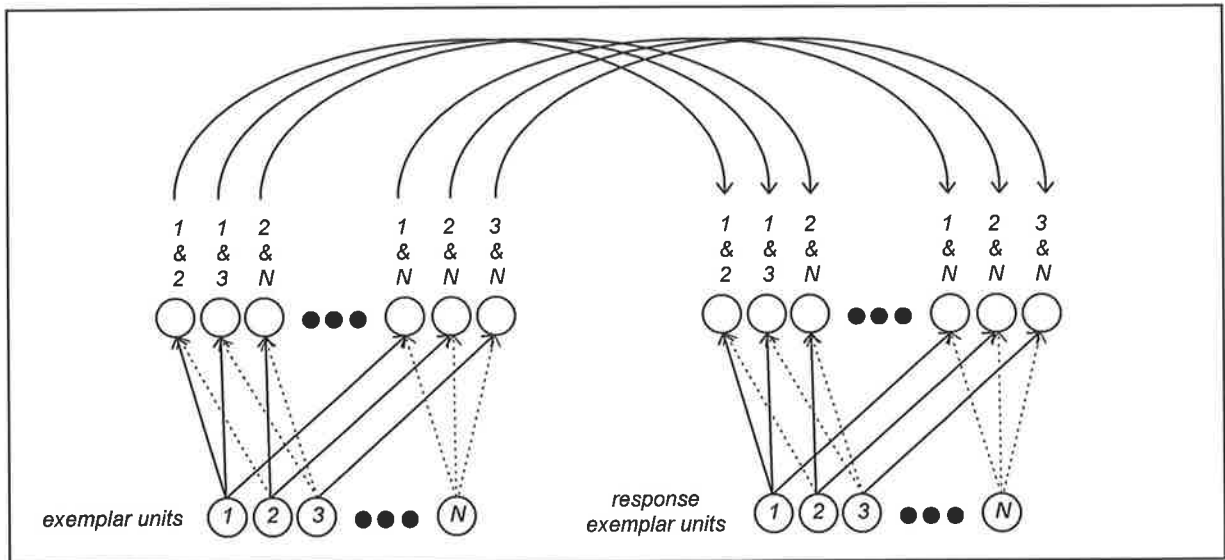
Of particular relevance here, however, is the application of the step-function sigmoid to allow only the topological properties of response space to constrain the psychological spaces derived by the model. There is certainly no difficulty introducing units with sigmoidal activation functions to the model, since this is the most widely employed activation function within connectionist modelling. Similarly, the setting of the  $\alpha$  parameter to a large value is unproblematic. Indeed, one motivation for the incorporation of sigmoidal activation functions is to approximate the step functions used in Perceptrons (Rosenblatt 1958), the ADALINE (Widrow & Hoff 1960), and other early ‘neural’ networks with a functional form which was differentiable.

A preliminary formulation of the way in which these units could be used to extend to the current model is depicted in Figure 9.3. The radial basis function architecture of both the psychological and response spaces has been appended with a layer of step function sigmoidal units. There is one such unit for each pair of stimuli, each taking as its input the difference between the similarities of these two stimuli, as generated across the exemplar and response exemplar layers. Thus, the sigmoidal units act to compare the relative similarity of each pair of stimuli to the current stimulus, both in terms of the current psychological space representation, and in terms of the environmental constraints contained in response space.

For example, the sigmoidal unit which compares stimuli 1 and 2 in either space has an excitatory connection of +1 to the similarity of stimulus 1 and an inhibitory connection of -1 to the similarity of stimulus 2, with the net input correspondingly being the difference between the two similarity values. Therefore, with reference to Figure 9.2, the derived activation of the sigmoidal unit will be either a fixed positive value if the difference between the similarities is positive (ie. if stimulus 1 is more similar to the presented stimulus than stimulus 2), or a fixed negative value if the difference between the similarities is negative (ie. if stimulus 1 is less similar to



the presented stimulus than stimulus 2). The fundamental feature of these activation values is that they do not contain any information regarding the magnitude of the difference between the similarity values. Rather, the sigmoid layers encode the topological ordering of the exemplar and response exemplar units. For each pair of stimuli, the value of the corresponding sigmoid unit indicates which of the two is closer to the currently presented stimulus.



*Figure 9.3. The envisaged extension to the current model, in which additional sigmoid layers are added to perform pairwise ordinal similarity comparisons of the exemplar and response exemplar activations values.*

Having extended the model in this way, the similarity error component of the internal error is appropriately reformulated as measuring the sum of the squared difference between the corresponding psychological and response space sigmoid units' activation values, rather than comparing the 'raw' similarity values. This form of error function is closely related to that employed by non-metric multidimensional scaling algorithms, since it only extracts ordinal level information from available target similarity measures. The gradient descent learning rule arising from the new internal error measure would operate to reposition stimulus points, seeking a configuration in which the ordering of the exemplars in psychological space matched the associated ordering of response exemplars in response space.

Clearly, under this scheme, the form of the basis functions used to generate the similarity measures in the exemplar and response exemplar units are largely irrelevant. Any monotonically decreasing basis function will produce the same pattern of activation across the introduced sigmoid layers. In particular, whilst the Universal Law of Generalisation implies that the final psychological space representational structure will relate distance to similarity with an exponential decay functional form, this structure will not be influenced by the form of the response space basis function.

#### 9.1.5. Dimensional Error

Section 5.2.3 concluded with the observation that the form of dimensional error given by

Equation 4.10 is worthy of further investigation. The method by which the error measure was derived, as documented in Section 4.2.2, hinges upon the notion of removing stimulus dimensions in such a way as to interfere minimally with the representational dictates of the similarity error. This motivation, in turn, suggested two qualitative features be required of the dimensional error form - that it increase as representational contribution increases, and that the rate of this increase should asymptotically approach zero. Whilst the functional form chosen meets these requirements, and the dimensional error measure ultimately employed in both the multidimensional scaling and mental representation learning models appears to be effective, there remains considerable scope for the establishment of more detailed theoretical foundations.

Progress in this area might be made through placing additional constraints, which embody reasonable assumptions regarding psychological space representations, upon the form of the dimensional error function. One source of such constraints derives from Bayesian considerations of optimal 'regularising' or 'penalty' functions within connectionist models (MacKay 1992a, 1992b). These functions, such as the weight decay term employed by Hinton (1989, see Section 2.3.1) and others, and the penalty term developed by Weigend, Rumelhart and Huberman (1991, see Equation 4.9) act to simplify or otherwise optimise the structure of a network. Clearly, the dimensional error measure fulfils precisely this sort of regularising role, since it effectively serves to remove units from the internal representation layer. As such, MacKay's (1992b) observation that:

"[a]lternative regularizers ... implicitly correspond to alternative hypotheses about the statistics of the environment" (p. 452)

suggests the development of a dimensional error measure which corresponds to the expected structure of psychological space representations.

For a given set of potential psychological space dimensions, there are perhaps two structural assumptions which could reasonably be made. First, the subset of potential dimensions which eventually constitute actual representational axes assume values drawn from a uniform distribution. This assumption follows directly from Shepard's (1987a) axiomatic postulation that each point in a psychological space is equally likely to be a stimulus point. Secondly, as argued in Section 4.3, each dimension in the subset of potential dimensions which are ultimately discarded from the derived representational structure assumes a (potentially) different constant value across all presented stimuli. The development of a regularising function which, in a Bayesian sense, optimally modified the connection matrix  $C$  in accordance with these two assumptions would seem to offer significant promise as a dimensional error measure.

In fact, two previously developed connectionist regularising functions reflect assumptions closely approximating those embodied by psychological spaces. First, Buntine and Weigend (1991) develop a learning rule which assumes connection weights are drawn from a mixture of a uniform distribution and a small variance zero-mean Gaussian distribution. The net effect of this

combination of distributions is neatly summarised by Zemel (1995):

“The Gaussian encourages small weight values to approach zero; the uniform distribution takes responsibility for larger weights and provides little pressure to change these values” (p. 574)

The incorporation of both a uniform distribution for representing information, and a Gaussian distribution for removing connections through weight decay, accords well with the representational requirements of psychological space. Even more encouraging is the explicit consideration of the interaction of the effects of these two distributions, which is reminiscent of the discussion in Section 4.2.2 from which desirable qualitative features of the dimensional error measure were derived. There are, however, several significant differences between the assumptions underlying Buntine and Weigend’s (1991) regularising function, and those required by psychological space representations.

Most fundamentally, the focus of Buntine and Weigend’s (1991) regulariser upon the removal of individual connection weights is not necessarily compatible with a focus upon the removal of the *sets* of connection weights which comprise individual stimulus dimensions, as required for psychological space dimensionality reduction. A potential stimulus dimension can only be removed when its entire ‘instar’ of connection weights have been relieved of the requirement to represent information. In addition, the use of a zero-mean Gaussian distribution is not entirely appropriate, since it removes connection weights by forcing the adoption of the specific value zero, rather than allowing an arbitrary fixed value. With reference to Figure 4.4, this lack of generality corresponds to collapsing the ‘line of non-contribution’ into the origin within conjugate psychological space.

The second regularising form, developed by Nowlan and Hinton (1992) fares significantly better in both of these regards. This approach involves fitting a mixture of Gaussian distributions to the connection weights. As noted by Nowlan and Hinton (1992, p. 476), since uniform distributions may be approximated by large variance Gaussian distributions, such a mixture is capable of approximating a mixture of uniform and Gaussian distributions, as required by psychological space representation. Furthermore, the limitations of zero-mean Gaussian distributions do not arise, since the various means of the distributions are modified during the course of learning. Even more importantly, the approach allows weights to be clustered “into subsets with the weights in each cluster having very similar values” (p. 473). By constraining these clusters of weights to correspond to the component dimensions of a psychological space, the regularising approach of Nowlan and Hinton (1992) offers considerable promise as the basis of a dimensional error measure. The development of these ideas should be a priority for future research.

There is a close relationship between Bayesian approaches to optimising the structure of a connectionist network, and approaches derived from complexity theory. Indeed, in some

circumstances, these two approaches give rise to formally equivalent conceptualisations of the optimality of connectionist models (Zemel 1995). It is interesting, therefore, to examine the implications of complexity theory for the models developed in this thesis.

The application of complexity theory to general connectionist modelling is most readily achieved through the ‘minimum description length’ approach (see Zemel 1995 for an overview). This approach formalises a measure which incorporates both the size, and the modelling accuracy, of a connectionist network. When optimised, this measure corresponds to the minimally complicated network which is capable of sufficient modelling accuracy. In effect, the minimum description length measure embodies the modelling principle variously known as the ‘Principle of Parsimony’ or ‘Ockham’s razor’, which insists that a model should only be as complicated as is necessary to account for the phenomena it seeks to explain and predict (see Casti 1992b, pp. 314-315 for an overview).

This tradeoff between accuracy and simplicity corresponds to the interaction of the similarity and dimensional errors in the connectionist multidimensional scaling and mental representation learning models. Indeed, given the embodiment of classical multidimensional scaling principles in these error measures, the minimum description length formalism may be applied directly to psychological space representations. The similarity error measures relates to the Universal Law of Generalization, seeking a representational configuration in which the distances between the stimulus points are appropriate. The dimensional error relates to the requirement that this representational configuration is of the minimum possible dimensionality. Psychological space representation as a whole, therefore, seeks to model similarity data in a way which is largely consistent with the minimum description length approach.

Indeed, under this view, the only way in which the fundamental principles of psychological space representation might be regarded as sub-optimal involves the strict adherence to geometric representation. Notions of algorithmic complexity can be taken to suggest that “the point of a scientific theory is to reduce the arbitrariness in the data” (Casti 1992b, p. 315), in the sense of providing a more compact means of reproducing the data. The psychological space theory of human mental representations certainly involves the compacting of information. For a set of  $N$  stimuli, a psychological space representation converts  $N \times N$  inter-stimulus similarity measures into  $N \times P$  coordinate locations, where  $P$  is the final dimensionality of the derived space. Since  $P$  is typically significantly smaller than  $N$ , this conversion constitutes a reduction in the resources required to represent the data. Furthermore, the invariant function specified by the Universal Law of Generalization provides a particularly efficient means of generating the similarity matrix from the coordinate values.

It does, however, seem possible that the restrictions placed upon a representational structure by the requirement that it be geometrically interpretable may sometimes be inappropriate. For

example, the critique of psychological space representation provided by Tversky and Hutchinson (1986, recall Section 3.1.5) suggests that nearest-neighbour constraints imposed by the nature of geometric representation do not allow for an optimal coding of some similarity matrices. It may well be the case that, for some stimulus sets, the optimal description of their similarity relationships, as formalised within complexity theory, also happens to be amenable to the geometric representation specified by the psychological space theory. For other stimulus sets, however, it is possible that these two theories imply different abstract representational structures. In this case, from the perspective of cognitive modelling, the advantage of psychological space theory - that of providing a graphically interpretable structure - seems less important than the ideal of cognitive economy promoted by complexity theory. An exploration of the relationship between these two approaches, and the possibility of developing mental representational structures based on notions of coding efficiency, appears to be a worthwhile undertaking.

---

## 9.2. Learning Psychophysical Mappings

As was noted in Section 4.1.1, the adoption of a local coding scheme at the stimulus input layer is theoretically well motivated, given the dangers inherent in representational pre-abstraction, but does limit the ways in which stimuli can be presented to a model. Both the multidimensional scaling and mental representation learning model adhere to this scheme, and are consequently are restricted to operating in stimulus domains which consist of a fixed and pre-determined stimulus set. Ultimately, however, the stimulation provided by the external world cannot be accommodated within a set structure. The continuous variation possible in sensory information is fundamentally incompatible with the discrete nature of sets. To be able to accommodate such stimulus domains, a coding scheme must be employed at the stimulus input layer which allows for continuous activation values and distributed stimulus representation.

With stimulus information being provided to a model in this way, the development of a psychological space representational structure requires the learning of a 'psychophysical mapping' between the stimulus input and internal representation layers. This mapping, in effect, transforms primitive sensory stimulation into abstract mental representation. As Daugman and Downing (1995) summarise:

"The external world presents itself only as physical signals at the sensory surface, which explicitly expresses very little of the information required for intelligent interaction with the environment. These signals must be converted into ... representations whose manipulation allows the organism or machine to bring an appropriate model of its external environment into contact with its external goals and purposes" (p. 414)

It is, therefore, of considerable importance to examine the possibility of extending the current model to operate in continuous stimulus domains through learning psychophysical

mappings. This possibility is, however, best approached through first examining the way in which a similar extension might be made to the multidimensional scaling model upon which the mental representation learning model was built.

### 9.2.1. Multidimensional Scaling Model

The only feedback available to the connectionist multidimensional scaling is that present in the pre-determined similarity matrix. Therefore, the model is only capable of directly learning with regard to those stimuli about which the matrix contains information. It is, however, possible to adapt the coding scheme employed at the stimulus input layer to allow an arbitrary number of stimuli to be presented to the model. As such, the possibility arises of extending the model to learn a psychophysical mapping of sorts, which allows the model to generate interpolated psychological space locations for novel stimuli.

For example, considering the colour stimulus domain examined in Section 5.1.1, the stimulus input layer could be reduced to a single unit, the activation value of which corresponded to the defining wavelength of the presented colour stimulus. The model could then learn psychological space locations for each of the fourteen particular wavelengths employed by Ekman (1954), as detailed in the similarity matrix given in Table 5.1. In so doing, a psychophysical mapping would be established between the stimulus input layer, where the stimulus is described in terms of a physical measurement of wavelength, and the internal representation layer, where the stimulus is described in terms of its psychological space location. The advantage of developing such a mapping, as discussed in relation to the model developed by Rumelhart and Todd (1993, refer Section 3.1.1), is that the model has the ability to generalise from the fourteen learned locations to derive appropriate psychological space locations for other stimuli with different wavelengths. Effectively, the model learns a continuous mapping from sensory to psychological description, in the form of an interpolative function, using the finite number of stimulus points about which it receives information.

Clearly, therefore, the ability of the extended model to develop psychophysical mappings resides in the ability of the network to accommodate appropriate functional mappings between the stimulus input and internal representation layers. Exactly what forms these mappings must be able to assume is difficult to determine. Merely assuming continuity might seem sufficiently general to encompass the functional forms which ultimately prove to be appropriate, although the characterisation of psychophysical mappings espoused by Gregson (1988, 1992, 1995) lies beyond these bounds. There are however, many stimulus domains for which assumptions such as continuity and smoothness would appear to suffice. For example, the colour stimulus domain involves a psychophysical mapping from a one-dimensional wavelength spectrum into a two-dimensional 'horseshoe' or 'colour circle' (recall Figure 3.1) which seems likely to meet both of

these assumptions.

In such cases, the function approximation capabilities of the radial basis function connectionist architecture (Hertz, Krogh & Palmer 1991, pp. 142-143, Poggio & Girosi 1990) are well suited to the modelling task. Consequently, the envisaged extension of the connectionist multidimensional scaling model to learn psychophysical mappings, as depicted in Figure 9.4, inserts a 'psychophysical mapping' layer which shares a radial basis function linkage with the stimulus input layer. This modification means, in principle, that any real-valued continuous mapping from a compact sensory domain into a psychological space may be accommodated by the model (Hartman, Keeler & Kowalski 1990).

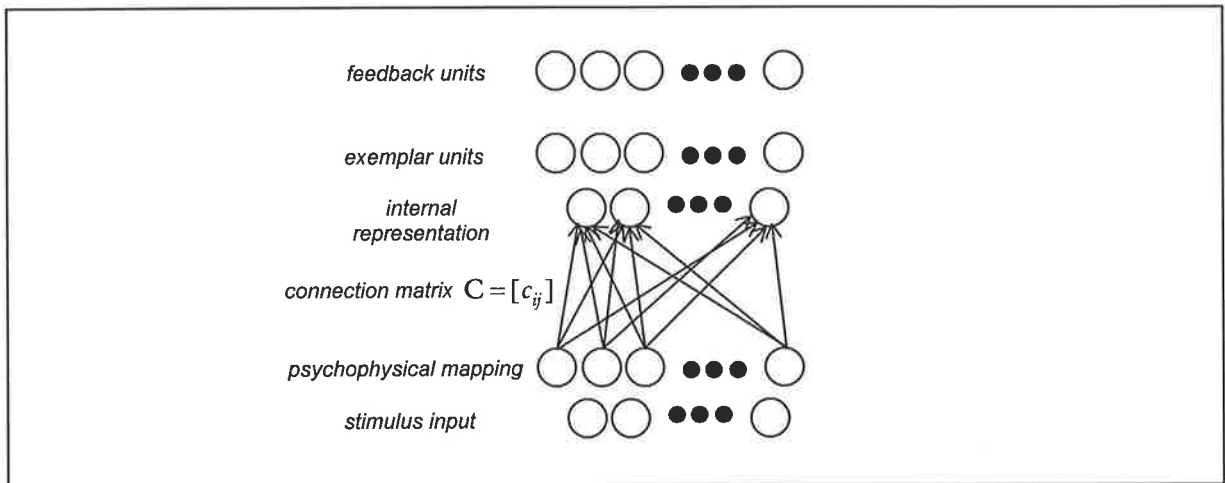


Figure 9.4. Envisaged psychophysical extension of the connectionist multidimensional scaling model.

The architectural extension of the model necessitates the development of a new learning rule. The total error measure employed in the original model remains entirely appropriate, and there is no reason not to continue to apply a gradient descent optimisation approach. The calculation of the required partial derivative, however, is made considerably more difficult by the adoption of a distributed representational scheme at the stimulus input layer. In particular, the effects of modifying the connection matrix  $C$  upon the locations of the non-presented stimuli in psychological space are not easy to determine. Nevertheless, given the information available from the exemplar units' locations, this difficulty does not seem insurmountable, possibly through the use of some suitable approximation. It is worth noting, however, that the pairwise stimulus presentation method employed in Rumelhart and Todd's (1993, recall Figure 3.9) model is particularly convenient in this regard.

Having adopted the radial basis function architecture, the possibility arises of incorporating an additional learning rule to move the psychophysical mapping units. Using the well established self-organisation approach (Kohonen 1982, 1984, 1990, Moody & Darken 1989, cf. Section 3.2.5), these units could be moved in such a way as to approximate the density distribution of stimulation being presented at the stimulus input layer. Thus, for example, if the model encountered coloured stimuli with wavelengths in a restricted range of the visible spectrum, the psychophysical units

would be adjusted to lie within this range. Through this prudent application of network resources, the 'resolution' with which the psychophysical map may be learned is improved. An even more promising possibility in this regard involves the introduction and deletion of psychophysical units during the model's operation as required. Using automatic complexity determination techniques such as those described by McMichael (1995), the covering map of psychophysical units might be able to accommodate a mapping of some desired resolution using a minimal amount of network resources.

A more significant impact of the connectionist multidimensional scaling model's extension to learn psychophysical mappings involves the placement of additional constraints upon the dimensional error measure. As discussed in Section 9.1.5, the form of the dimensional error employed in both models is somewhat arbitrary, and requires the development of additional constraints for further refinement. Whilst the sources derived from Bayesian and complexity considerations remain relevant, the interpolative form dictated by psychophysical mappings could also guide the construction of the dimensional error measure. One of the impacts of regularising terms, such as the dimensional error, relates to the manner in which a continuous mapping is interpolated from a set of learned point-to-point mappings (see, for example, Poggio & Girosi 1990). Since the extended model constructs psychophysical mappings in precisely this way, assumptions regarding the appropriate form of these mappings, as derived from more general psychophysical theory, are appropriately embedded within the dimensional error measure.

### **9.2.2. Mental Representation Learning Model**

In one sense, the psychophysical extension of the multidimensional scaling model is equally applicable to the mental representation learning model. Architecturally, as is evident from Figures 4.1 and 7.1, the two models have identically connected stimulus input and internal representation layers. Consequently, as is shown in Figure 9.5, a psychophysical mapping layer could also be incorporated into the mental representation learning model. Suitable learning rules for the psychophysically extended multidimensional scaling model could also readily be adapted to extend the mental representation learning model.

This approach, however, does not fully realise the potential for psychophysically extending the mental representation learning model. In particular, the capability of the mental representation learning model to encounter a complicated external environment is significantly under-utilised. The multidimensional scaling model, through its canonical re-interpretation of classic techniques, relies on a similarity matrix for the provision of information. This restriction makes the assumption that the feedback received by the model is derived from a discrete matrix structure unavoidable. It is, however, inappropriate for a psychophysical extension of the mental representation learning model to maintain this practice.



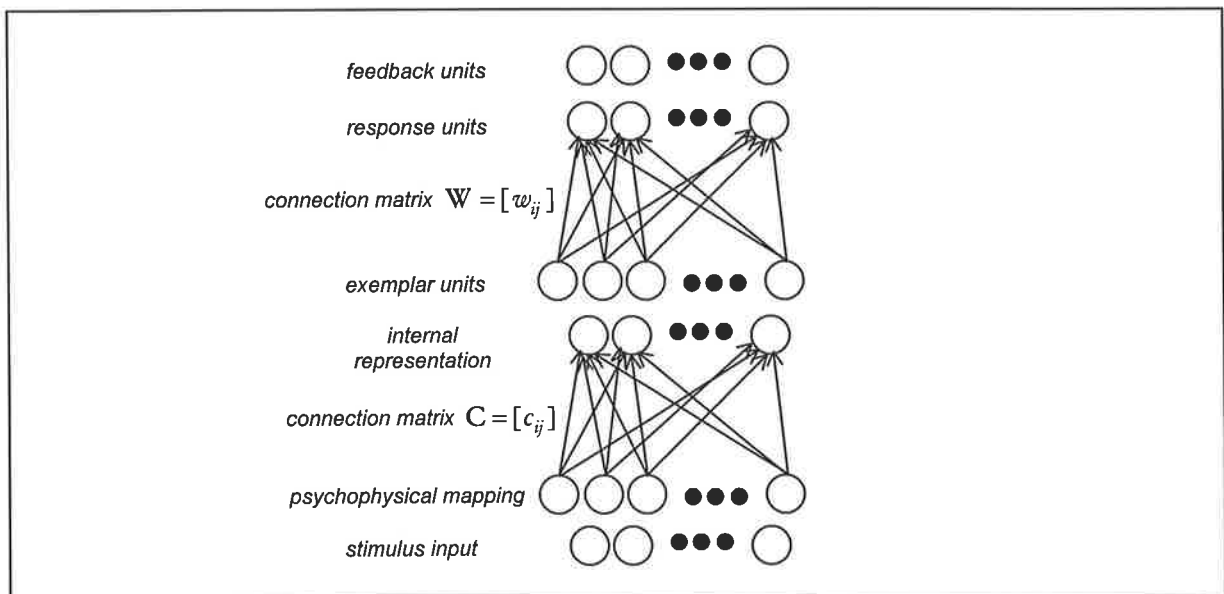


Figure 9.5. Possible psychophysical extension of the mental representation learning model.

A corollary of the primary motivation for developing psychophysical extensions - to allow the models to encounter continuous stimulus domains - is that the environmental feedback be made available for arbitrary stimuli. In other words, the way in which the environment is modelled should have the sophistication to generate an appropriate pattern of sensory properties and categorical associations for any of presented stimulus selected from a continuous stimulus domain. As was discussed in Section 6.2.1, the realism with which the encompassing environment is simulated directly influences the model's claim to be a model of the learning of mental representations. Allowing every stimulus, rather than a set of pre-determined stimuli, in a continuous domain to be encountered constitutes a significant advancement in this regard. Unfortunately, the way in which this might be achieved in the mental representation learning model is far from clear.

Clearly, the adoption of continuous stimulus domains is incompatible with the way in which the exemplar and response exemplar layers are constructed, since the units in these layers have a one-to-one association with elements of a stimulus set. In one sense this difficulty is not insurmountable, since a 'covering map' of units in both the psychological and response spaces could be introduced. In this way, the mental representational learning model would be able to approximate the information content of both its psychological and response spaces without reliance on a pre-determined stimulus set. Effectively, such an extension is analogous to that involved in the ALCOVE model's extension of the ALEX model (recall Section 3.2.5). The difficulty arises, however, in attempting to maintain the crucial relationship between the psychological and response spaces. As was emphasised in Section 7.3.1, the response space acts as the information source from the which the representational structure in psychological space is derived. The means by which the necessary exchange of information takes place is through the canonical alignment of the units in exemplar and response exemplar layers. Specifically, since each

stimulus in the environmental domain corresponds to a pair of units - one in the response exemplar layer and one in the exemplar layer - the former unit appropriately provides information to the latter. The similarity error component of the internal error measure directly reflects this relationship in its comparison of the activation of associated exemplar and response exemplar units (recall Equation 7.9).

Within a covering map representational scheme, however, such canonical alignment is not available. There is no way of determining the way in which the information contained in response space should constrain the representations derived in psychological space on the basis of common invariant relationships to environmental stimuli. Instead, each point in response space is properly associated with a continuous multivariable function which indicates the psychological similarity between that point and every other point in the space. The learning of an appropriate psychophysical mapping, therefore, involves developing a psychological space representational structure in which these similarity functions are accommodated in as few dimensions as possible. Although there is some possibility such constraints might be able to be implemented using interpolative approximation techniques, the way in which the relationships in the general case would change over time, and operate in spaces of potentially changing and different dimension, makes the true psychophysical extension of the mental representation learning model a difficult topic for future research.

## Chapter 10: Concluding Remarks

Having introduced, developed, evaluated, and discussed a range of possible extensions to both the mental representation learning model, and the multidimensional scaling model upon which it was based, it would seem appropriate to conclude by reflecting upon the general cognitive modelling principles set forth in Chapter 1. In particular, it is interesting to measure the degree to which the purported strengths of the connectionist framework have been utilised in the development of the final mental representation learning model.

The model is certainly consistent with the notions of situated and embodied agency. The entire emphasis of the model's development from its connectionist multidimensional scaling base is one based on environmentally constraining the mental representations it learns. Indeed, the model's ability to learn principled psychological space structures through a process of general cognitive interaction with an external world is potentially one of its greatest strengths. It should be restated, however, that the presented demonstrations of the model fall well short of this ultimate goal. As was acknowledged in Section 6.2.3., the simulated environments provided to the model contain neither the complexity nor the inherent dynamism of the real world. What they do provide is a manageable and manipulable means of examining the learning properties of the model, particularly in relation to its basis in environmentally constrained mental representation. The introduction of successively more sophisticated simulated environments, culminating in the application of real world constraints through genuine physical situatedness and embodiment, constitutes a natural and worthwhile extension to the current model.

With regard to the principle of modelling mental representations as emergent cognitive phenomena, the success of the mental representation learning model is more difficult to gauge. Certainly, the psychological spaces learned by the model arise from a complicated interaction with response space. From a non-linear dynamical perspective, the model's learning may be viewed in terms of an attempt to structure its psychological space to 'resonate' with the environmental information captured in response space. The final mental structures developed by the model, under this conception, correspond to the attractors defined by the sensory properties and categorical associations of the stimulus set, whilst the continually changing environment serves to periodically reshape the surrounding basins of attraction. This type of interaction, as highlighted in relation to the notion of equilibration (recall Section 7.4.2), seems broadly consistent with the notion of emergent mental structures.

It is less clear, however, that the final mental representations themselves adhere equally well to this principle. In particular, the extent to which the mental representations are "statistically emergent active symbols" (Hofstadter 1985, p. 659) could validly be questioned, since the specification of a point in a coordinate space as the representation of a stimulus does not seem to

confer any inherent capability of activity. Any deficiency of the model in this regard, of course, amounts to a shortcoming of the psychological space theory which postulates this form of mental representational structure. The criticism of psychological space representations, from the emergent cognition perspective, is that which Hofstadter (1985) directed at symbolic approaches to cognitive modelling, in asserting that they:

“bypass epiphenomena (“collective phenomena” if you prefer) by simply installing structures that mimic the superficial features of those epiphenomena” (Hofstadter 1985, p. 642)

It is not the case that psychological spaces are symbolic cognitive models, as is evident from the ease of their connectionist interpretation (recall Section 3.2.3) as distributed representational structures. What is true, however, is that the adoption of the radial basis function approach negates the possibility of psychological spaces being sub-conceptually modelled through their underlying consequential regions. Indeed, this limitation was explicitly discussed in Section 3.2.6, where it was suggested that the consequential region modelling approach afforded no additional flexibility which was required to model the acquisition of human mental representation. For example, the generalisation gradient which emerges from the action of consequential regions was considered, on the basis of Shepard’s (1987a) derivation, to be sufficiently well approximated by the Universal Law of Generalization. To the extent that the mental representation learning model fulfils its cognitive modelling goal, this assumption is well grounded, and notions of modelling parsimony continue to take preference over those of emergent cognition.

It remains, however, entirely possible, if not likely, that the extension of the mental representation learning model to model a more comprehensive range of cognitive phenomena might require the adoption of the sub-conceptual consequential region approach. For example, as was mentioned in Section 3.2.6, chronometric considerations do not seem readily addressable within the radial basis function approach. Indeed, simply attempting to accurately model the time course of mental representation learning, as canvassed in Section 9.1, might require more flexibility than is currently available. Whilst, in such situations, it is tempting to impose the newly required capabilities upon the established psychological space structure, this practice has its limitations. Gregson (1995) is emphatic in this regard, asserting that such an over-extension of cognitive process modelling constitutes:

“an evasive argument by some writers committed to the notion of an underlying psychological space. Instead of trying to find a new model of such a space ... observed ... data are then postulated to be in some way biased by attention and memory weighting variables, which proliferate prodigiously” (p. 206)

It is difficult to specify a precise point at which psychological spaces, as incorporated in the mental representation learning model, become overly rigid and static characterisations of those realised by the ‘first-principles’ modelling of consequential regions. This boundary between

parsimony and flexibility is both inherently subtle, and dependent upon the cognitive modelling goals at hand. What can be acknowledged, however, is that the mental representation learning model developed in this thesis is limited by its deliberate failure to incorporate fully the principle of sub-conceptually modelling mental representations as emergent phenomena. Future sufficiently ambitious modelling goals might require the model to be rebuilt, rather than extended.

## REFERENCE LIST

- Ackley, D.H., Hinton, G.E. & Sejnowski, T.J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9, 147-169.
- Anderson, J.A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.
- 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 (3), 409-429.
- Anderson, J.R. & Schooler, L.J. (1991). Reflections of the environment in memory. *Psychological Science*, 2 (6), 396-408.
- Anderson, J.R. (1992). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14, 471-517.
- Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika*, 56 (4), 567-587.
- Ash, T. & Cottrell, G. (1995). Topology-modifying neural network algorithms. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 990-993). Cambridge, MA: MIT Press.
- Ashby, F.G. & Maddox, T.W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372-400.
- Ashby, F.G. & Perrin, N.A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, 95 (1), 124-150.
- Averill, E.W. (1993). Hidden kind classification. *Psychology of Learning and Motivation*, 29, 437-467.
- Barsalou, L.W. (1983). Ad hoc categories. *Memory & Cognition*, 11, 211-227.
- Barsalou, L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 629-654.
- Baxter, J. (1996). *The canonical distortion measure for vector quantization and approximation*. Unpublished manuscript, Department of Mathematics, London School of Economics and Department of Computer Science, Royal Holloway, University of London.
- Bezdek, J.C. & Pal, N.R. (1995a). A note on self-organizing semantic maps. *IEEE Transactions on Neural Networks*, 6 (5), 1029-1036.
- Bezdek, J.C. & Pal, N.R. (1995b). An index of topological preservation for feature extraction. *Pattern Recognition*, 28 (3), 381-391.
- Borg, I. (1982). Scaling - A review of German scaling literature of the last fifteen years. *The German Journal of Psychology*, 7 (1), 63-79.
- Borg, I. & Lingoes, J. (1987). *Multidimensional similarity structure analysis*. New York: Springer Verlag.
- Bourne, L.E. (1974). An inference model of conceptual rule learning. In R.L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 231-256). Potomac, MD: Erlbaum.
- Brooks, R.A. (1991a). Intelligence without representation. *Artificial Intelligence*, 47, 139-159.
- Brooks, R.A. (1991b). Intelligence without reason. *Proceedings of the twelfth international conference on AI*, 569-595.
- Buntine, W.R. & Weigend, A.S. (1991). Bayesian back-propagation. *Complex Systems*, 3 (2), 603-643.
- Carlton, E.H. & Shepard, R.N. (1990a). Psychologically simple motions as geodesic paths I. Asymmetric objects. *Journal of Mathematical Psychology*, 34, 127-188.
- Carlton, E.H. & Shepard, R.N. (1990b). Psychologically simple motions as geodesic paths II. Symmetric objects. *Journal of Mathematical Psychology*, 34, 189-228.
- Carpenter, G.A. (1989). Neural network models for pattern recognition and associative memory. *Neural Networks*, 2, 243-257.
- Carpenter, G.A. & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organising neural network, *Computer*, 21 (3), 77-88.

- Carpenter, G.A., Grossberg, S. & Reynolds, J.H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 4, 565-588.
- Casti, J.L. (1989). *Paradigms lost: Images of man in the mirror of science*. London: Sphere Books.
- Casti, J.L. (1992a). *Reality rules: I*. New York: Wiley.
- Casti, J.L. (1992b). *Reality rules: II*. New York: Wiley.
- Chalmers, D.J., French, R.M. & Hofstadter, D.R. (1991). *High-level perception, representation, and analogy: A critique of Artificial Intelligence methods*. CRCC Technical Report 49. Bloomington, IN: Indiana University.
- Choi, S., McDaniel, M.A. & Busemeyer, J.R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition*, 21 (4), 413-423.
- Collins, A.M. & Loftus, E.F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82 (6), 407-428.
- Coombs, C.H. (1958). An application of a nonmetric of multidimensional analysis of similarities. *Psychological Reports*, 4, 511-518.
- Corter, J.E. & Gluck, M.A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111 (2), 291-303.
- Cox, T.F. & Cox, M.A.A. (1991). Multidimensional scaling on a sphere. *Communications in Statistics: Theory and Methods*, 20 (9), 2943-2953.
- Cox, T.F. & Cox, M.A.A. (1994). *Multidimensional scaling*. London: Chapman & Hall.
- Cussins, A. (1990). The connectionist construction of concepts. In M.A. Boden (Ed.), *The philosophy of artificial intelligence* (pp. 368-441). Oxford: Oxford University Press.
- Cutting, J.E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7 (1), 71-87.
- Cutting, J.E., Proffitt, D.R. & Kozlowski, L.T. (1978). A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4 (3), 357-372.
- Daugman, J. & Downing, C. (1995). Gabor wavelets for statistical pattern recognition. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 414-419). Cambridge, MA: MIT Press.
- Dyer, M.G. (1988). The promise and problems of connectionism. *Behavioral and Brain Sciences*, 11, 32.
- Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, 38, 467-474.
- Ennis, D.M. (1988a). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, 117 (4), 408-411.
- Ennis, D.M. (1988b). Toward a universal law of generalization. *Science*, 242, 944.
- Ennis, D.M. (1992). Modeling similarity and identification when there are momentary fluctuations in psychological magnitudes. In F.G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 279-298). Hillsdale, NJ: Erlbaum.
- Fillenbaum, S. & Rappaport, A. (1971). *Structure in the subjective lexicon*. New York: Academic Press.
- Flexer, A. (1996). *Limitations of self-organizing maps for vector quantization and multidimensional scaling*. Unpublished manuscript, Austrian Research Institute for Artificial Intelligence.
- French, R.M. (1991). *Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks*. CRCC Technical Report 51. Bloomington, IN: Indiana University.
- Freyd, J.J. (1987). Dynamic mental representations. *Psychological Review*, 94 (4), 427-438.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition tolerant of deformation and shifts in position. *Biological Cybernetics*, 36 (4), 193-202.
- Fukushima, K. (1988). A neural network for visual pattern recognition. *Computer*, 21 (3), 65-75.
- Garner, W.R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.
- Garner, W.R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

- Garner, W.R. (1993). How the mind works, if there is one. In G. Harman (Ed.), *Conceptions of the mind: Essays in honor of George A. Miller* (pp. 161-171). Hillsdale, NJ: Erlbaum.
- Gati, I. & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 8 (2), 325-340.
- Getty, D.J., Swets, J.A., Swets, J.B. & Green, D.M. (1979). On the prediction of confusion matrices from similarity judgements. *Perception & Psychophysics*, 26, 1-19.
- Gibson, J.J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton-Mifflin.
- Gibson, J.J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.
- Glenberg, A.M. (in press). What memory is for. *Behavioral and Brain Sciences*.
- Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52, 125-157.
- Goodhill, G.J., Simmen, M.W. & Willshaw, D.J. (1994). *An evaluation of the use of multidimensional scaling for understanding brain connectivity*. Centre for Cognitive Science, Research Paper EUCCS/RP-63. Edinburgh, UK: University of Edinburgh.
- Grau, J.W. & Nelson, D.K. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General*, 117(4), 347-370.
- Gregson, R.A.M. (1975). *Psychometrics of similarity*. New York: Academic Press.
- Gregson, R.A.M. (1988). *Nonlinear psychophysical dynamics*. Hillsdale, NJ: Erlbaum.
- Gregson, R.A.M. (1992). *N-dimensional nonlinear psychophysics*. Hillsdale, NJ: Erlbaum.
- Gregson, R.A.M. (1995). *Cascades and fields in perceptual psychophysics*. Singapore: World Scientific.
- Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Boston, MA: Reidel.
- Handel, S. (1967). Classification and similarity of multidimensional stimuli. *Perceptual and Motor Skills*, 24, 1191-1203.
- Hanson, S.J. & Burr, D.J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, 13, 471-518.
- Hartman, E.J., Keeler, J.D. & Kowalski, J.M. (1990). Layered neural networks with Gaussian units as universal approximations. *Neural Computation*, 2, 210-215.
- Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Englewood Cliffs, NJ: Macmillan.
- Hertz, J., Krogh, A. & Palmer, R.G. (1991). *Introduction to the theory of neural computing*. Redwood City, CA: Addison-Wesley.
- Hilgard, E.R. & Bower, G.H. (1975). *Theories of learning*. Englewood Cliffs, NJ: Prentice-Hall.
- Hinton, G.E. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40, 185-234.
- Hinton, G.E., McClelland, J.L. & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1* (pp. 77-109). Cambridge, MA: MIT Press.
- Hinton, G.E. & Sejnowski, T.J. (1983). Optimal perceptual inference. *Proceedings of IEEE Computer Science Conference on Computer Vision and Pattern Recognition, 1983*, 448-453.
- Hofmann, T. & Buhmann, J. (1994). Multidimensional scaling and data clustering. In J.D. Cowan, G. Tesauro and J. Alspector (Eds.), *Advances in neural information processing systems 7* (pp. 459-466). San Mateo, CA: Morgan Kaufman.
- Hofstadter, D.R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. New York: Basic Books.
- Hofstadter, D.R. (1985). *Metamagical themes: Questing for the essence of mind and pattern*. New York: Basic Books.
- Hofstadter, D.R. (1988). Foreword. In P. Kanerva, *Sparse Distributed Memory* (pp. xi-xviii). Cambridge, MA: MIT Press.



- Hubert, L., Arabie, P. & Hesson-Mcinnis, M. (1992). Multidimensional scaling in the city-block metric: A combinatorial approach. *Journal of Classification*, 9, 211-236.
- Humphrey, N. (1992). *A history of the mind*. Vintage Books.
- Intrator, N. & Edelman, S. (1996). Making a low-dimensional representation suitable for diverse tasks. *Connection Science*, 8 (2), 205-223.
- Kac, M. & Ulam, S.M. (1968). *Mathematics and logic*. Penguin.
- Klock, H. & Buhmann, J. (1997). Multidimensional scaling by deterministic annealing. To appear in *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, Venice, 1997.
- Knapp, A.G. & Anderson, J.A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10, 616-637.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.
- Kohonen, T. (1984). *Self-organization and associative memory*. New York: Springer-Verlag.
- Kohonen, T. (1988a). An introduction to neural computing. *Neural Networks*, 1, 3-16.
- Kohonen, T. (1988b). The 'neural' phonetic typewriter. *Computer*, 21 (3), 11-22.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78 (9), 1464-1480.
- Kolen, J.F. & Pollack, J.B. (1991). Backpropagation is sensitive to initial conditions. In R.P. Lippmann, J.E. Moody & D.S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 860-867). San Mateo, CA: Morgan Kaufman.
- Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, 112 (3), 500-526.
- Krantz, D.H. (1967). Rational distance functions for multidimensional scaling. *Journal of Mathematical Psychology*, 4, 226-245.
- Kruschke, J.K. (1990). *A connectionist model of category learning*. Unpublished doctoral dissertation: University of California at Berkeley.
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99 (1), 22-44.
- Kruschke, J.K. (1993a). Three principles for models of category learning. *The Psychology of Learning and Motivation*, 29, 57-90.
- Kruschke, J.K. (1993b). Human category learning: Implications for backpropagation models. *Connection Science*, 5, 3-36.
- Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29 (1), 1-27
- Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29 (2), 28-42.
- Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.
- Lawrence, S., Tsoi, A.C. & Back A.D. (1996). Function approximation with neural networks and local methods: Bias variance and smoothness. In P. Bartlett, A. Burkitt and R.C. Williamson (Eds.), *Proceedings of the seventh Australian conference on neural networks* (pp. 16-21). Canberra, ACT: Australian National University.
- LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. & Jackel, L.D. (1990). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 541-551.
- Leithold, L. (1986). *The calculus with analytic geometry*. New York: Harper & Row.
- Lengellé, R. & Denœux, T. (1996). Training MLPs layer by layer using an objective function for internal representation. *Neural Networks*, 9 (1), 83-97.
- Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review*, 102 (1), 185-191.

- Lewis, M.D. (1994). Reconciling stage and specificity in neo-Piagetian theory: Self-organizing conceptual structures. *Human Development*, 37 (3), 143-169.
- Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA: MIT Press.
- Lindman, H. & Caelli, T. (1978). Constant curvature Riemannian scaling. *Journal of Mathematical Psychology*, 17, 89-109.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21 (3), 105-117.
- Lowe, D. (1995). Radial basis function networks. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 779-782). Cambridge, MA: MIT Press.
- Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush and E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: John Wiley and Sons.
- Lynch, G.S., Granger, R.H., Larson, J. & Baudry, M. (1989). Cortical encoding of memory: Hypotheses derived from analysis and simulation of physiological learning rules in anatomical structures. In L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish (Eds.), *Neural connections, mental computation* (pp. 180-224). Cambridge, MA: MIT Press.
- MacKay, D.J.C. (1992a). Bayesian interpolation. *Neural Computation*, 4, 415-447.
- MacKay, D.J.C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4, 448-472.
- Markman, A.B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General*, 118 (4), 417-421.
- McCarthy, J. (1986). Epistemological challenges for connectionism. *Behavioral and Brain Sciences*, 11, 44.
- McClelland, J.L., O'Reilly, R.C. & McNaughton, B.L. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102 (3), 419-457.
- McClelland, J.L. & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, 375-407.
- McClelland, J.L. & Rumelhart, D.E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2. Cambridge, MA: MIT Press.
- McMichael, D.W. (1995). Automatic complexity determination of Gaussian mixture models with the EMS algorithm. In P. Bartlett, A. Burkitt and R.C. Williamson (Eds.), *Proceedings of the Seventh Australian Conference on Neural Networks*, (pp. 103-108). Canberra, ACT: Australian National University.
- Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist*, 44 (12), 1469-1481.
- Medin, D.L., Altom, M.W., Edelson, S.M. & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 37-50.
- Medin, D.L. & Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117 (1), 68-85.
- Medin, D.L. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). New York: Cambridge University Press.
- Medin, D.L. & Schaffer, M.M. (1978). Context theory of classification. *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.
- Mervis, C.B. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89-115.
- Merz, C.J. & Murphy, P.M. (1996). *UCI repository of machine learning databases* [<http://www.ics.uci.edu/~mlern/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
- Miller, G.A. (1956). The magical number seven, plus or minus two. *Psychological Review*, 63, 81-97.
- Miller, G.A. (1962). *Psychology: The science of mental life*. Harmondsworth, Middlesex: Penguin.

- Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.
- Minsky, M. (1986). *The society of mind*. New York: Simon & Schuster.
- Miyano, H. & Inukai, Y. (1982). Sequential estimation in multidimensional scaling. *Psychometrika*, 47 (3), 321-336.
- Moody, J. & Darken, C.J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1, 281-294.
- Myung, I.J. (1994). Maximum entropy interpretation of decision bound and context models of categorization. *Journal of Mathematical Psychology*, 38, 1-31.
- Neisser, U. (1987). From direct perception to conceptual structure. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 11-24). Cambridge, MA: Cambridge University Press.
- Newell, A. & Simon, H.A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the Association for Computing Machinery*, 19, 113-126.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4, 135-183.
- Noble, M.E. & Bahrick, H.P. (1956). Response generalization as a function of intratrask response similarity. *Journal of Experimental Psychology*, 51 (6), 405-412.
- Norman, D.A. (1993). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science*, 17, 1-6.
- Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10 (1), 104-114.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115 (1), 39-57.
- Nosofsky, R.M. (1988a). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14 (1), 54-65.
- Nosofsky, R.M. (1988b). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14 (4), 700-708.
- Nosofsky, R.M. (1988c). On exemplar-based representations: Reply to Ennis (1988). *Journal of Experimental Psychology: General*, 117 (4), 412-414.
- Nosofsky, R.M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, 2 (6), 416-421.
- Nosofsky, R.M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43, 25-53.
- Nosofsky, R.M. & Kruschke, J.K. (1992). Investigations of an exemplar-based connectionist model of category learning. *The Psychology of Learning and Motivation*, 28, 207-250.
- Nosofsky, R.M., Kruschke, J.K. & McKinley, S.C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18 (2), 211-233.
- Nowlan, S.J. & Hinton, G.E. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4, 473-493.
- Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101 (4), 608-631.
- Palmer, S.E. (1991). Goodness, Gestalt, groups, and Garner. In J.R. Pomerantz and G.L. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 23-39). Washington, DC: American Psychological Association.
- Piaget, J. (1970). Piaget's theory. In P.H. Mussen (Ed.), *Carmichael's manual of child psychology, vol. 1* (pp. 703-732). New York: Wiley.

- Pipkin, J.S. (1982). Some remarks on multidimensional scaling in geography. In R.G. Golledge and J.N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets* (pp. 214-232). Minneapolis, MN: University of Minnesota Press.
- Poggio, T. & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78 (9), 1481-1497.
- Posner, M.I. & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353-363.
- Posner, M.I. & Keele, S.W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304-308.
- Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97 (2), 285-308.
- Reed, R. & Marks, R.J. (1995). Neurosmithing: Improving neural network learning. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 639-644). Cambridge, MA: MIT Press.
- Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black and W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). New York, NY: Cambridge University Press.
- Ritter, H. (1990). Self-organising maps for internal representations. *Psychological Research*, 52 (2/3), 128-136.
- Rosch, E. (1978). Principles of categorization. In E. Rosch and B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-77). Hillsdale, NJ: Erlbaum.
- 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.
- Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386-408.
- Rumelhart, D.E. (1980). Schemata: The building block of cognition. In R. Spiro, B. Bruce and W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.
- Rumelhart, D.E. (1989). The architecture of mind: A connectionist approach. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 133-159). Cambridge, MA: MIT Press.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533-536.
- Rumelhart, D.E. & McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89 (1), 60-94.
- Rumelhart, D.E., Smolensky, P., McClelland, J.L. & Hinton, G.E. (1986). Schemata and sequential thought processes in PDP models. In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2* (pp. 7-57). Cambridge, MA: MIT Press.
- Rumelhart, D.E. & McClelland, J.L. (1986). On learning the past tenses of English verbs. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1* (pp. 216-271). Cambridge, MA: MIT Press.
- Rumelhart, D.E. & Todd, P.M. (1993). Learning and connectionist representations. In D.E. Meyer and S. Kornblum (Eds.), *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 2-37). Cambridge, MA: MIT Press.
- Russell, S.J. & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Sejnowski, T.J. & Rosenberg, C.R. (1987). Parallel networks that learn to produce English text. *Complex Systems*, 1, 145-168.
- Schank, R.C. & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.
- Schiffman, S.S., Reynolds, M.L. & Young, F.W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.

- Schneider, R.B. (1992). A uniform approach to multidimensional scaling. *Journal of Classification*, 9, 257-273.
- Shanks, D.R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17 (3), 433-443.
- Shanks, D.R. & Gluck, M.A. (1994). Tests of an adaptive network model for the identification and categorization of continuous-dimension stimuli. *Connection Science*, 6 (1), 59-89.
- Shannon, C.E. & Weaver, W. (1949). *The mathematical theory of communication*. Chicago: University of Illinois Press.
- Shepard, R.N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22 (4), 325-345.
- Shepard, R.N. (1958a). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 55 (6), 509-523.
- Shepard, R.N. (1958b). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, 65(4), 242-256.
- Shepard, R.N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27 (2), 125-140.
- Shepard, R.N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27 (3), 219-246.
- Shepard, R.N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D.I. Mostofsky (Ed.), *Stimulus generalization* (pp. 94-110). Stanford, CA: Stanford University Press.
- Shepard, R.N. (1972). Psychological representation of speech sounds. In E.E. David and P.B. Denes (Eds.), *Human communication: A unified view* (pp. 67-113). New York: McGraw Hill.
- Shepard, R.N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39 (4), 373-422.
- Shepard, R.N. (1975). Form, formation, and transformation of internal representations. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 87-122). Potomac, MD: Erlbaum.
- Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210, 390-398.
- Shepard, R.N. (1981a). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology*, 24, 21-57.
- Shepard, R.N. (1981b). Psychophysical complementarity. In M. Kubovy and J.R. Pomerantz (Eds.), *Perceptual organization* (pp. 279-341). Hillsdale, NJ: Erlbaum.
- Shepard, R.N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89 (4), 305-333.
- Shepard, R.N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, 91 (4), 417-447.
- Shepard, R.N. (1987a). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Shepard, R.N. (1987b). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 251-275). Cambridge, MA: MIT Press.
- Shepard, R.N. (1988a). Time and distance in generalization and discrimination: Reply to Ennis (1988). *Journal of Experimental Psychology: General*, 117 (4), 415-416.
- Shepard, R.N. (1988b). George Miller's data and the development of methods for representing cognitive structures. In W. Hirst (Ed.), *The making of cognitive science* (pp. 45-70). Cambridge, MA: Cambridge University Press.
- Shepard, R.N. (1988c). Toward a Universal Law of Generalization [Response]. *Science*, 242, 944.

- Shepard, R.N. (1989). Internal representation of universal regularities: A challenge for connectionism. In L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish (Eds.), *Neural connections, mental computation* (pp. 104-134). Cambridge, MA: MIT Press.
- Shepard, R.N. (1990a). Neural nets for generalization and classification: Comment on Staddon and Reid (1990). *Psychological Review*, 97 (4), 579-580.
- Shepard, R.N. (1990b). *Mind sights*. New York: W.H. Freeman and Company.
- Shepard, R.N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J.R. Pomerantz and G.L. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53-71). Washington, DC: American Psychological Association.
- Shepard, R.N. (1992a). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In J.H. Barkow, L. Cosmides and J. Tooby, (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 495-532). New York: Oxford University Press.
- Shepard, R.N. (1992b). The advent and continuing influence of mathematical learning theory: Comment on Estes and Burke. *Journal of Experimental Psychology: General*, 121(4), 419-421.
- Shepard, R.N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, 1(1), 2-28.
- Shepard, R.N. & Arabie, P. (1979). Additive clustering: Representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86, 87-123.
- Shepard, R.N. & Cooper, L.A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.
- Shepard, R.N., Hovland, C.L. & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75 (13), Whole No. 517.
- Shepard, R.N. & Kannappan, S. (1991). Connectionist implementation of a theory of generalization. In R.P. Lippmann, J.E. Moody and D.S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 665-671). San Mateo, CA: Morgan Kaufman.
- Shepard, R.N. & Tenenbaum, J. (1991). *Connectionist modeling of multidimensional generalization*. Paper presented at the meeting of the Psychonomic Society, San Francisco, November 1991.
- Simon, H.A. (1981). *The sciences of the artificial*. Cambridge, MA: MIT Press.
- Simon, H.A. & Kaplan, C.A. (1989). Foundations of cognitive science. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 1-47). Cambridge, MA: MIT Press.
- Sjoberg, L. & Thorslund, C. (1979). A classificatory theory of similarity. *Psychological Research*, 40 (3), 223-247.
- Smith, E.E. (1989). Concepts and induction. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 501-526). Cambridge, MA: MIT Press.
- Smolensky, P. (1988a). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11, 1-23.
- Smolensky, P. (1988b). Putting together connectionism - again. *Behavioral and Brain Sciences*, 11, 59-74.
- Staddon, J.E.R. & Reid, A.K. (1990). On the dynamics of generalization. *Psychological Review*, 97 (4), 576-578.
- Tenenbaum, J.B. (1996). Learning the structure of similarity. To appear in D.S. Touretzky, M.C. Mozer and M.E. Hasselmo (Eds.), *Neural information processing systems 8*. Cambridge, MA: MIT Press.
- Todd, P.M. & Rumelhart, D.E. (1995). *Feature abstraction from similarity ratings: A connectionist approach*. Unpublished manuscript, Stanford University.
- Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84 (4), 327-352.
- Tversky, A. & Gati, I. (1982). Similarity, separability and the triangle inequality. *Psychological Review*, 89 (2), 123-154.
- Tversky, A. & Hutchinson, J.W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93 (1), 3-22.

- van der Maas, H.L.J. & Molenaar, P.C.M. (1992). Stagemwise cognitive development: An application of catastrophe theory. *Psychological Review*, 99 (3), 395-417.
- Vera, A.H. & Simon, H.A. (1993). Situated action: A symbolic interpretation. *Cognitive Science*, 17, 7-48.
- Vickers, D. (1979). *Decision processes in visual perception*. London: Academic Press.
- Vickers, D. (1996). An Erlanger programme for psychology. Paper presented at *Symposium on Chaos Theory Applied to Social Science Research, Fourth International Social Science Methodology Conference*, Colchester, Essex, U.K., July 1996.
- Vickers, D., Vincent, N. & Medvedev, A. (1996). The geometric structure, construction, and interpretation of path-following (trail-making) tests. *Journal of Clinical Psychology*, 52 (6), 651-661.
- Wallach, M.A. (1958). On psychological similarity. *Psychological Review*, 65 (2), 103-116.
- Webb, A.R. (1995). Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition*, 28 (5), 753-759.
- Weigend, A.S., Rumelhart, D.E. & Huberman, B.A. (1991). Generalization by weight-elimination with application to forecasting. In R.P. Lippmann, J.E. Moody and D.S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 875-882). San Mateo, CA: Morgan Kauffman.
- Widrow, B. & Hoff, M.E. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record* (pp. 96-104). New York: IRE.
- Zemel, R.S. (1995). Minimum description length analysis. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 572-575). Cambridge, MA: MIT Press.