



ORBIT - Online Repository of Birkbeck Institutional Theses

Enabling Open Access to Birkbeck's Research Degree output

Attentional control in categorisation: towards a computational synthesis

<https://eprints.bbk.ac.uk/id/eprint/40486/>

Version: Full Version

Citation: He, Liusha (2020) Attentional control in categorisation: towards a computational synthesis. [Thesis] (Unpublished)

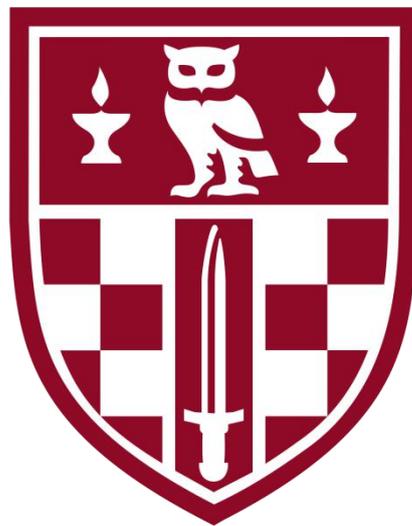
© 2020 The Author(s)

All material available through ORBIT is protected by intellectual property law, including copyright law.

Any use made of the contents should comply with the relevant law.

[Deposit Guide](#)
Contact: [email](#)

Attentional Control in Categorisation: Towards a Computational Synthesis



Liusha He

Department of Psychological Sciences,
Birkbeck, University of London

This thesis is submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy (Ph.D.) at Birkbeck, University of London

Abstract

This thesis develops an integrated computational model of task switching in heterogeneous categorisation by combining theories of cognitive control and category learning. The thesis considers the strengths and shortcomings of a range of existing computational accounts of categorisation (ALCOVE, SUSTAIN, ATRIUM and COVIS) by reimplementing each and applying each to human data from the categorisation literature. It is argued that most of these models cannot account for heterogeneous categorisation, i.e., situations where the category structure includes subsets with incompatible boundaries. Moreover, the only one of the four computational models that can account for heterogeneous categorisation, ATRIUM, does not completely account for the influence of top-down control during categorisation tasks. The models are also limited because they are based purely on feedforward principles, and while they are able to learn to categorise stimuli adequately, they do not account for categorisation response times, or for task-switching effects observed in recent research on heterogeneous categorisation. In order to address these limitations, the thesis presents a model that combines an interactive activation account of task-switching with a modular architecture of categorisation. The model is shown to successfully simulate reaction time costs and effects of preparation time on task switching.

Declaration

I declare that the work presented in this thesis is my own. Where it builds on other people's work or ideas this is clearly marked.

Table of Contents

| | |
|---|------------|
| Chapter 1. General Introduction..... | 1 |
| 1.1 Thesis Motivation..... | 1 |
| 1.2 Thesis Outline..... | 3 |
| Chapter 2. Human Categorisation: Empirical and Computational Approaches..... | 5 |
| 2.1 Introduction..... | 5 |
| 2.2 Multiple Representations In Category Learning: Computational Approaches..... | 6 |
| 2.3 Implied Cognitive Control In Category Learning..... | 18 |
| 2.4 Concluding Comments..... | 30 |
| Chapter 3. Attention to Categorisation: The Perspective from Exemplar Theory..... | 32 |
| 3.1 Introduction..... | 32 |
| 3.2 Reimplementing ALCOVE..... | 34 |
| 3.3 Theoretical Implications..... | 40 |
| Chapter 4. Case Study I: SUSTAIN and Multiple Representations..... | 45 |
| 4.1 Introduction..... | 45 |
| 4.2 Reimplementing SUSTAIN..... | 47 |
| 4.3 Simulation Studies..... | 52 |
| 4.4 Theoretical Implications..... | 66 |
| Chapter 5.69Case Study II: Multiple Systems Theory and Modular Architecture of COVIS..... | 69 |
| 5.1 Introduction..... | 69 |
| 5.2 COVIS..... | 71 |
| 5.3 Simulation Study: Hybrid Category Learning..... | 79 |
| 5.4 Theoretical Implications..... | 86 |
| Chapter 6. Case Study III: The Modular Architecture of ATRIUM and the Mechanism of Representational Attention..... | 90 |
| 6.1 Introduction..... | 90 |
| 6.2 Reimplementing ATRIUM..... | 93 |
| 6.3 Simulation Study: Hybrid Category Learning..... | 100 |
| 6.4 Representational Attention and Task Switching..... | 105 |
| 6.5 Theoretical Implications..... | 113 |
| Chapter 7. Attention to Multiple Representations: The Perspective from Supervisory System Theory..... | 118 |
| 7.1 Introduction..... | 118 |
| 7.2 Supervisory System: A Theory of Cognitive Control..... | 120 |
| 7.3 A Modular Model of Task Switching..... | 124 |
| 7.4 The Hyperdirect Pathway Hypothesis..... | 129 |

| | |
|---|------------|
| 7.5 Toward a New Modular Architecture of Human Categorisation..... | 131 |
| Chapter 8. CATHER: A Combined Model of Task Switching Effects in Categorisation..... | 136 |
| 8.1 Introduction..... | 136 |
| 8.2 Erickson’s (2008) Task Inhibition Proposal..... | 137 |
| 8.3 Model Formalisation..... | 140 |
| 8.4 Model Applications..... | 149 |
| 8.4.1 Application I: Erickson’s (2008) Task Switching Effects..... | 150 |
| 8.4.2 Application II: Helie’s (2017) Preparation Effects..... | 153 |
| 8.5 Discussion..... | 155 |
| Chapter 9. Reflection and Conclusion..... | 158 |
| 9.1 Summary..... | 158 |
| 9.2 Theoretical Implications..... | 159 |
| 9.3 Limitations and Future Directions..... | 164 |
| 9.4 Conclusions..... | 165 |
| Appendix A. A Preliminary Behavioural Study..... | 167 |
| A.1 Background..... | 167 |
| A.2 Method..... | 169 |
| A.3 Results and Discussion..... | 171 |
| Appendix B. Attention Learning Mechanism and Wisconsin Card Sorting Task..... | 176 |
| B.1 Introduction..... | 176 |
| B.2 The Wisconsin Card Sorting Task..... | 177 |
| B.3 The Model..... | 178 |
| B.4 Results and Discussion..... | 181 |
| Appendix C. Modelling Aha and Goldstone (1992)..... | 183 |
| C.1 Background..... | 183 |
| C.2 Method..... | 184 |
| C.3 Results and Discussion..... | 185 |
| References..... | 187 |

Chapter 1.

General Introduction

1.1 Thesis Motivation

Category learning, otherwise known as concept learning, is a fundamental aspect of human cognition. It enables appropriate behaviour by allowing category information to be used to guide behaviour rather than knowledge specific to individual objects, which may not be known. It is reasonable to assume multiple representations subserve category learning as multiple representations can facilitate performance on a variety of tasks in the daily life. In the past two decades, therefore, the primary focus of the computational approaches in the research of category learning were made in establishing that humans use multiple representations (e.g., Anderson & Betz, 2001; Ashby et al., 1998; Erickson & Kruschke, 1998; Love et al., 2004; Nosofsky et al., 1998).

Given that there are multiple representations that together produce a single categorisation response, a question to ask is if there exists any common mechanism that mediates selection between separate representations. In some situations where the task can be best addressed by a single representation, a simple competition mechanism is enough. For example, in Ashby et al.'s (1998) dual systems account, COVIS, it is assumed that a verbal system that selects and tests explicit rules, and a nonverbal system that learns implicit stimulus-response mappings, compete to determine the output responses. For a given stimulus, the system that can make the less

equivocal decision wins the competition, and over time, the model learns to favour the system that has won the most overall.

COVIS has received great success in accounting for dissociations between explicit and implicit category learning. However, there are some situations that require shifts between separate category representations. For example, some complex categorisation tasks can be simplified by decomposing them into separate local solutions. The heterogeneity of category representation is accommodated by a contextual framework called task partitioning (Erickson, 2008), also known as knowledge partitioning (Yang & Lewandowsky, 2003; 2004; Lewandowsky et al., 2006). Task partitioning in heterogeneous category learning is helpful for establishing a linkage between cognitive control and category learning, because, in addition to learning to select appropriate representations for all stimuli, in heterogeneous category learning, participants are required to switch from subtask to subtask on a trial-by-trial basis. Indeed, not surprisingly, task switching costs in heterogeneous category learning have been found in more recent research (Crossley et al., 2017; Erickson, 2008; Helie, 2017).

However, the primary focus of previous research was not on establishing an accurate model of how control is passed back and forth between multiple representations. Among all the existing computational approaches, feed-forward neural networks with an error-driven, back-propagation (BP) algorithm have become the mainstream (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Kruschke, 1992; 1996; 2001; Love et al., 2004; see Kruschke, 2011 for review), because these BP networks are well-suited for associative learning. But feed-forward networks may not be able to simulate task switching, and traditional research does not allow a more accurate model of control of multiple representations to be constructed. Based upon the research directed at task switching in heterogeneous category learning, the time now seems propitious to direct attention to incorporating a task switching mechanism into category learning. For building a new model, some significant revisions of existing computational approaches are required.

1.2 Thesis Outline

The main purpose of the present research is to develop an integrated model of task switching in heterogeneous categorisation by combining an account of cognitive control with existing category learning networks. For doing so, it is useful to first revisit some existing associative learning networks of categorisation, because the most straightforward approach for reaching the integration purpose is presumably to incorporate together the principles of task switching and category learning.

In Chapter 2, the main recent computational and empirical approaches in the domain of research of human categorisation are introduced. For empirical approaches, the focus is specifically on evidence that suggests the importance of cognitive control in category learning. Whereas, for computational approaches, four network models, namely, ALCOVE, SUSTAIN, ATRIUM and COVIS, are briefly reviewed. Although ALCOVE is not a multiple representations model, it remains meaningful for theoretical development because ALCOVE is the most classic network model of category learning, and the other three models can all be somehow considered as extensions or variants of the exemplar-based model.

In Chapters 3, 4, 5 and 6, the four network models, ALCOVE, SUSTAIN, COVIS and ATRIUM, respectively, are revisited in depth. In each case, a reimplementations of the model is described and tested against existing data. The strengths and shortcomings of each case and their implications for theoretical development are then discussed. SUSTAIN and COVIS are excluded from further analyses, because SUSTAIN lacks stability in learning continuous dimensions, whereas COVIS does not involve cognitive control in switching between multiple representations. In contrast, although the gating network in ATRIUM which suggests a stimulus-dependent representation (Erickson, 2008; Erickson & Kruschke, 1998) in categorisation may be unrealistic, its modular architecture remains meaningful for task switching.

In Chapter 7, an influential cognitive control theory, the Supervisory System theory, is introduced, as is a computational model of task switching derived from this theory, the Gilbert-Shallice (2002) task switching model. The modular architectures of the Gilbert-Shallice model and the ATRIUM model share some similarities. Therefore, in Chapter 8, a new modular

network model that combines exemplar theory and task switching theory is developed. This model receives support in that it is shown to successfully simulate switch costs in reaction time and effects of preparation time on task switching.

All simulations reported in the thesis were conducted in MatLab, All source code is available at:

Github: https://github.com/LiushaHe0317/Thesis_source_code

Open science: <https://osf.io/mpy3w/>

Chapter 2.

Human Categorisation: Empirical and Computational Approaches

2.1 Introduction

Making sense of the environment is a major computational challenge for human beings. Fortunately, the cognitive capacity of categorisation allows us to decrease the computational load by treating novel situations in accord with existing category representations or rules, rather than as situations for which little is known. However, the rules guiding categorisation behaviour in daily life are diverse. For some tasks, such as ‘stop at a red light’, as they are easy to be expressed verbally, we can quickly capture the rules. However, for some other tasks, such as distinguishing between musical instruments, instead of using verbal rules, a person may need to achieve the task based on the overall similarity. Given the diversity of categorisation tasks in the real world, it is reasonable to have multiple representations to mediate the complex tasks and processes.

Although there remains some disagreement (e.g., Craig & Lewandowsky, 2012; Lewandowsky et al., 2011; Newell et al., 2010), a body of research in categorisation has suggested that there are multiple representations involved (Erickson & Kruschke, 1998; Love et al., 2004; Nosofsky et al., 1998), and forming of these representations are associated with distinct neural systems (Ashby et al., 1998). For example, Ashby et al. (1998) proposed that verbal rule-based categorisation tasks are primarily associated with activities in frontal areas, whereas performance on similarity-based categorisation tasks are related to the functioning of the basal

ganglia (BG). Likewise, Kruschke and Erickson (1994) suggested that there are two types of representations in category learning, abstract rule-based representation and exemplar-based representation.

Given the divergence of category learning representations, an important question would be how we regulate these representations to generate appropriate single behaviour in response to the dynamic world. In the past decade, this interesting question has attracted the attention of more and more researchers. A number of valuable studies have been conducted. It is, thus, necessary to systematically review the more recent work.

The aim of this chapter is to review the empirical evidence and theoretical developments in the domain of multiple category representations. This review will consist of three sections. In the first section, the traditional empirical and computational approaches used to support the framework of multiple representations in category learning will be summarised. The existing network models based on multiple representations will be included. In the second section, some recent research on the relationship between cognitive control and category learning will be reviewed. Evidence from research on categorisation automaticity and the interaction between multiple category learning systems is discussed. The third section focuses on recent research concerning heterogeneous category learning and task switching, and the implications for developing a new account of cognitive control in category learning are discussed.

2.2 Multiple Representations In Category Learning: Computational Approaches

2.2.1 Different Categorisation Task Types

The relationship between category structure and task difficulty has long been of interest to cognitive scientists. In the most classic study, known as the six types of categorisation tasks (the logical structure of the six problems is shown in Fig 2-1 (Left)), Shepard, Hovland and Jenkins (1961) found that the order of difficulty, measured in error rates, could be predicted by the number of dimensions required to specify the category membership (see also Kruschke, 1993).

When the number of relevant dimensions is equal in two different tasks, one task might be easier than another if a portion of the stimuli in that task can be classified using a sub-optimal strategy (see also Lewandosky, 2011; Love et al., 2004; Nosofsky et al., 1994a for more details). Nosofsky et al (1994a) replicated and extended the Shepard et al (1961) study. Fig 2-1 (Right) shows the proportion of incorrect responses as a function of block, as observed by Nosofsky et al (1994).

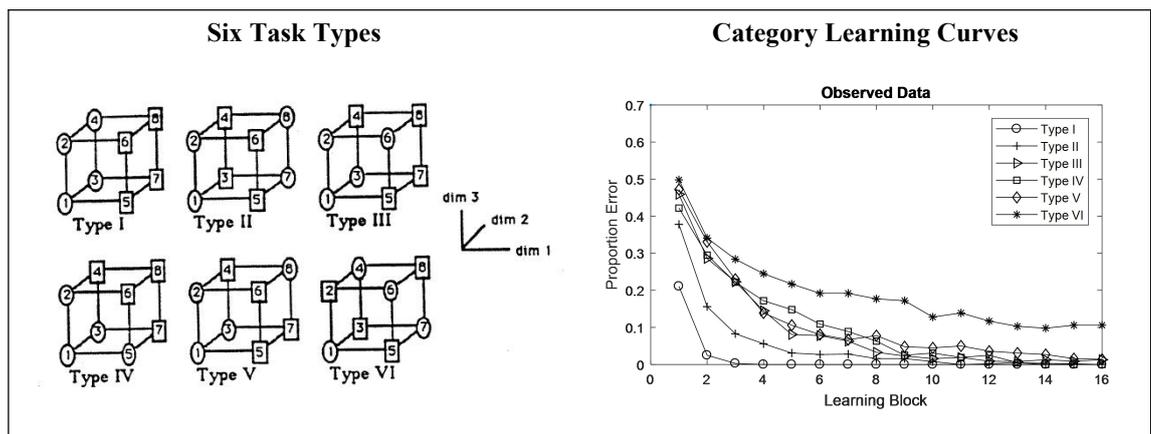


Fig 2-1. Left Panel: A schematic illustration of six types of classification problems examined by Shepard, Hovland, and Jenkins (1961) (dim = dimension). For each problem type, items in one category are indicated by squares and those in the other category are shown by circles; Right Panel: Observed human learning data from Nosofsky et al.'s (1994) replication of Shepard et al.'s (1961) experiment.

In fact, the Shepard et al.'s (1961) six types of categorisation task types can be further divided into four categories: namely, 1) one-dimensional (1D) rule-based categorisation task (Type I); 2) two-dimensional (XOR, i.e., exclusive-or) rule-based categorisation task (Type II); 3) multidimensional-structured categorisation tasks (Types III, IV and V); and 4) unstructured categorisation task (Type VI). It is noteworthy that, in addition to the characteristic of involving fewer relevant dimensions than the other tasks, the rules that maximise accuracy of 1D and XOR tasks are easy to describe verbally. From a computational perspective, these rules can be represented using IF/THEN rules (i.e., 'if dim 1 is 1, then respond A, else respond B'). The categorisation tasks determined by verbal rules (e.g., Types I and II), in the recent literature, are referred to as rule-based (RB) tasks.

However, the rules that maximise accuracy of the multidimensional structured categorisation tasks appear to be difficult to describe verbally (Anderson & Betz, 2001; Ashby et al., 1998; Nosofsky et al., 1994). Accuracy of these task types can be maximised either by

integrating information from multiple dimensions at some predecisional stage — an information-integration (II) strategy (Ashby et al., 2003; Ashby & Gott, 1988) — or by a rule-plus-exception strategy, meaning the case in which most of the category stimuli are categorised according to an explicit rule but some stimuli are exceptions to the rule, and participants must shift between strategies (e.g., Denton et al., 2008; Erickson & Kruschke, 1998; Love et al., 2004; Nosofsky et al., 1994a). Consequently, tasks that are not easily described by verbal rules are referred to, in some cases, as information-integration (II) tasks, but in some other cases, they are termed rule-plus-exception tasks. Real-world examples of II tasks are common. For example, deciding whether an x-ray shows a tumour not only involves some explicit reasoning but also shares many properties with II-type tasks. Years of training are required, and expert radiologists are only partially successful at describing their categorisation strategies. Examples of rule-plus-exception tasks in the daily life include past tense inflection of verbs where most verbs are regular (-ed), while a few verbs are exceptions (e.g., gave, made, grew, etc.). Note that the representations used by an II strategy and a rule-plus-exception strategy are completely different in nature. However, as the number of exemplars is limited, one problem of Shepard et al.'s paradigm is that the II and rule-plus-exception strategies are not easily distinguished in task Types III, IV and V.

2.2.2 Exemplar Theory and Attention Learning Networks

Nosofsky (1986) proposed the exemplar theory and a computational account called the generalised context model (GCM). According to GCM, people represent categories by storing individual exemplars in memory, and classify objects based on their similarity to these stored exemplars. To model similarity relations among exemplars, GCM adopts a so-called 'multidimensional scaling' approach in which exemplars are represented as points in a multidimensional psychological space. More importantly, Nosofsky's (1986) exemplar theory assumes that the term similarity is highly context-dependent, and this context-dependent nature of similarity should be determined by the dimensional attention strengths that systematically modify the structure of the psychological space in which the exemplars are embedded. Thus,

dimensional attention strengths serve to ‘stretch’ the psychological space along highly attended, relevant dimensions, but to ‘shrink’ the space along unattended, irrelevant dimensions. In the language of Nosofsky (2011), the stretching and shrinking effects of dimensional attention a ‘have profound influence on similarity relations among exemplars and on the resulting classification predictions from the model’.

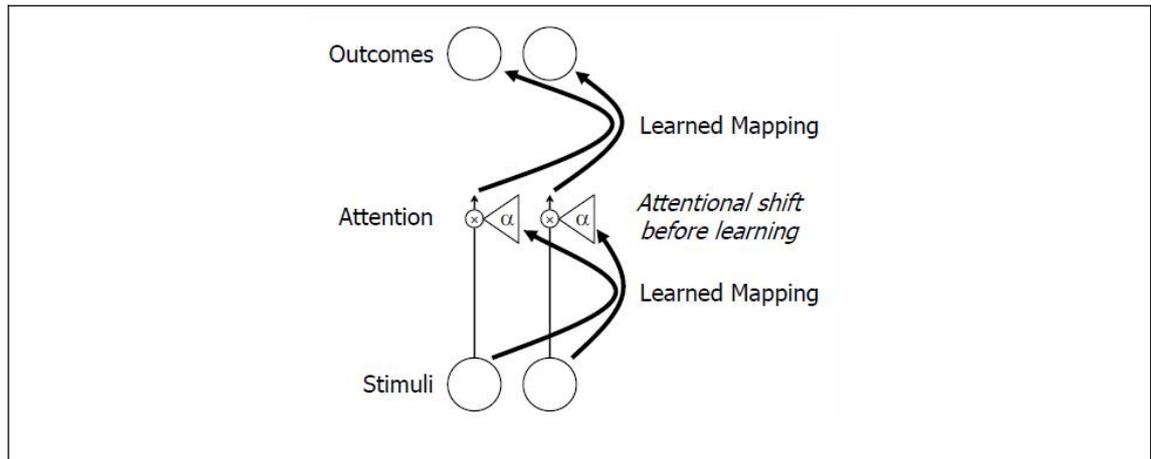


Fig 2-2. General framework of feed-forward network models for attention learning. The stimuli are represented at the bottom of the diagram by activations of corresponding nodes. Thick curved arrows denote learned associative mappings. Attention is denoted by “ α ” in the middle layer and acts as a multiplier on the stimuli. On a given trial of learning, attention is shifted before the mappings are learned.

GCM has received great success in categorisation, and there have been numerous extensions of the original exemplar-based account. One of the highly influential extension is Kruschke’s (1992) network model, ALCOVE (attention learning covering map). The ALCOVE model embeds the exemplar-based representation of GCM within a framework of a feed-forward, attention learning network (see Fig 2-2) (Kruschke, 2011). The dimensional attention and associative memory strengths on individual exemplars are no longer hand-set parameters, but are learned by an error reduction mechanism. In addition, following the general framework of attention learning networks, ALCOVE can be adapted to different specific variants, depending on situations to be modelled and the complexity demanded by the data. Successors of ALCOVE include RASHNL (rapid attention shifts and learning, Kruschke and Johansen, 1999), ADIT (attention to distinctive input, Kruschke, 1996a), EXIT (extension of ADIT, Kruschke, 2001) and so on. The general framework of attention learning networks is able to account for several

complex phenomena, such as highlighting (Kruschke, 1996a; 2001) and blocking (Kruschke & Blair, 2000) in category learning.

Now, consider again the Shepard paradigm. Of course, it is not a problem for the attention learning network to fit the learning data of different task types (Kruschke, 1992; Love et al., 2004; Nosofsky et al., 1994a), but when simulating learning in tasks of Type III, IV and V, does the network fairly implement the form of representation in the human mind? ALCOVE is more likely to take an II strategy and not a rule-plus-exception strategy. This is because, although the attention learning model assume that the processes of category learning should involve the process to store the exemplars in memory, it, in fact, does not implement it in that way. Instead, the exemplar nodes, in the applications, are predefined as the nodes in the hidden layer. This issue limits the attention learning network to account for the multiple representation situations, as it only produces a single representation for a single categorisation task.

Love et al. (2004), thus, proposed a new attention learning network, SUSTAIN (supervised and unsupervised stratified adaptive incremental network), that implements (in their view) how humans learn categories from exemplars. The principles of category learning in SUSTAIN can be simply summarised as: a) simple first, SUSTAIN starts with simple rules; b) stimuli that share similarities are clustered together in memory (prototype effects); c) attentional selection is competitive. Like ALCOVE, SUSTAIN also involves an attention learning mechanism, in which attention strength on each dimension changes according to task difficulty. The smaller the difference among the learned dimensional attention strengths, the more substructures are needed. This mechanism may facilitate the use of a rule-plus-exception strategy. But, unlike ALCOVE, as an incremental network, exemplars in SUSTAIN are not predefined, but recruited or redefined according to a ‘surprising event’ principle. To be more specific, there are no hidden nodes predefined in the network; the first exemplar recruited and stored in the network is the stimulus presented in the first trial. The membership of subsequent stimuli is determined by the similarity between each stimulus and the most activated exemplar. If there is any item that owns more typicality than the first exemplar, then it redefines the most activated exemplar. If an error (a surprising event) happens, a new exemplar is then recruited. In addition, SUSTAIN considers the exemplars as substructures of a category structure. Thus, it is assumed that, in the multiple representations situation, these substructures compete to determine the categorisation response.

SUSTAIN applies few very simple principles but it has been successfully fit to several data sets (see Love et al, 2004), including the Shepard paradigm.

2.2.3 Multiple Systems Theory of Category Learning

2.2.3.1 Dissociations Between Separate Category Learning Tasks

Although Shepard et al. (1961) provided a standard paradigm for subsequent category learning studies, the use of deterministic category structures (and some other binary-valued categorisation tasks, see also Medin et al., 1982; Medin & Schaffer, 1978, etc.) has obvious weaknesses (e.g., Alfonso-Reese et al., 2002; Ashby & Valentin, 2017; Kruschke, 1993; Love et al., 2004). Participants could solve the task Types III, IV and V by purely remembering the small number of training exemplars. In our daily life, the features (dimensions) of many objects are continuous and probabilistically distributed in the relevant psychological space, such as length, width, or spatial frequency of geometric figures and continuous changes in colours. The category structures based on continuous-valued features sometimes involve a large number of training exemplars, and, thus, avoid the use of unexpected strategies.

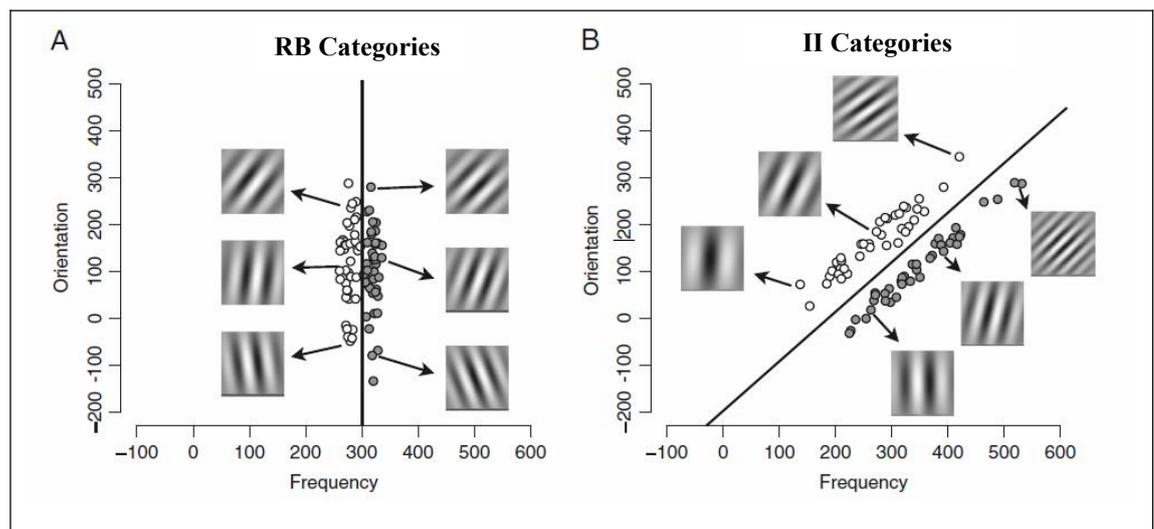


Fig 2-3. Examples of continuous-valued rule-based (A) and information-integration (B) stimuli and category structures.

Ashby and Gott (1988) introduced the general re-sampling randomisation technique (GRRT) by which, in a multi-dimensional, continuous psychological space, probabilistic category structures are defined by a multivariate normal distribution (e.g., Ashby et al., 2003; Ashby & Waldron, 1999; McKinley & Nosofsky, 1996; Miles et al., 2014; Zeithamova & Maddox, 2006). In large-size categorisation tasks, the optimal decision boundaries of RB tasks are orthogonal to coordinate axes (see Fig 2-3A), while optimal decision boundaries of II tasks are diagonal or at oblique angles to coordinate axes (see Fig 2-3B). By using the large-size category structure paradigm, it seems impossible to solve the II categorisation tasks by remembering the training exemplars or using the rule-plus-exception strategy.

Earlier investigations of Ashby and his colleagues have confirmed that, as an independent category representation, II category learning is dissociated from RB category learning in several ways. It has been confirmed that success in RB tasks strongly depends on working memory and executive attention, but previous research has not provided direct evidence to show a correlation between II tasks and working memory (Ashby & Maddox, 2010; Sloutsky, 2010). For example, RB tasks are disrupted by a dual working memory task much more than II tasks (e.g., Waldron & Ashby, 2001; Miles et al, 2014; Zeithamova & Maddox, 2006; 2007). In addition, reducing the reaction time (RTs) (Smith et al., 2015), levels of category overlap (Ell & Ashby, 2006), delayed or segmentary feedback (e.g., Ashby & O'Brien, 2007; Maddox et al., 2003; Maddox & Ing, 2005), and switching location of response keys, interferes much more in II tasks than RB tasks (Ashby et al., 2003; Nosofsky et al., 2005). In addition, substantial evidence from cognitive neuroscience suggests that learning of these two types of categorisation tasks depends on distinct neural circuits. For example, in a neuroimaging study reported by Nomura et al. (2007), participants were scanned while learning both RB and II categories. The results of RB category learning showed greater differential activation in the hippocampus, the anterior-cingulate cortex (ACC), and prefrontal cortex (PFC). At the same time, II categorisation exhibited greater differential activation in the striatum, especially the caudate nucleus (see also Cincotta & Seger, 2007). Moreover, a number of neuropsychological studies also confirmed that compared with undergraduate controls (young group) and aged-matched controls (mean age = 67.87, SD = 5.38), patients with basal ganglia dysfunctions, such as Parkinson's disease (PD) and Huntington's

disease (HD), were significantly impaired in the performance of RB categorisation tasks, but not II tasks (e.g., Ashby et al., 2003b; Ell et al., 2006; 2010).

It is noteworthy that, however, there remains some contrary evidence against this dissociation perspective. For example, Price et al. (2009) proposed that PD patients' performance on RB and II tasks is strongly affected by medication. They found that, although neurochemical changes associated with PD may disrupt rule shifting and selection, a typical treatment may cause problems in rule maintenance. In this sense, some deficits in category learning may depend crucially on whether the patients are tested 'on' or 'off' medication. But, in Ashby and colleagues' studies, this factor seems to be ignored. In addition, Nosofsky and Kruschke (2002) demonstrated that the exemplar-based model, ALCOVE, can naturally predict the behavioural pattern observed by Waldron and Ashby (2001). Furthermore, Newell et al. (2010) failed to replicate the selective dual task interference effects found by Zeithamova and Maddox (2006), which also challenged the dissociation perspective. In addition, a more recent study by Donkin et al. (2015) showed that the model-based strategy identification technique used by Ashby and colleagues is problematic. They found that by incorporating uncertainty in individuals' strategy identification, the proportion of II strategy users became significantly lower than when using standard practice. More evidence against Ashby's view will be given in Sections 2.3 and 2.4.

2.2.3.2 COVIS

Computational models have played a prominent role in shaping our understanding of human category learning. The neurocomputational account of multiple systems in category learning—COVIS (COmpetition between Verbal and Implicit Systems, Ashby et al., 1998; 2011) is one of the influential network models. COVIS postulates that there are two category learning systems acting in parallel that compete to control the action selection in category learning. One system is a frontal-based, explicit learning system and the other system is a striatal-based, procedural learning system. This idea of dual systems derives from the research on dissociations of RB and II categorisation tasks. As mentioned above, the strategies used in RB tasks can be commonly described verbally. In most common applications, only one stimulus dimension is relevant, and the participant's task is to discover this relevant dimension and then to map the different

dimensional values to the relevant categories (e.g., Fig 2-4A). The functions of the explicit learning system include selecting and testing a set of verbal rules. Tasks using RB category structures are learned quickly in the explicit learning system. A wide range of category structures are learned in the procedural learning system, which occurs in a reinforcement fashion and depends heavily on reliable and immediate feedback.

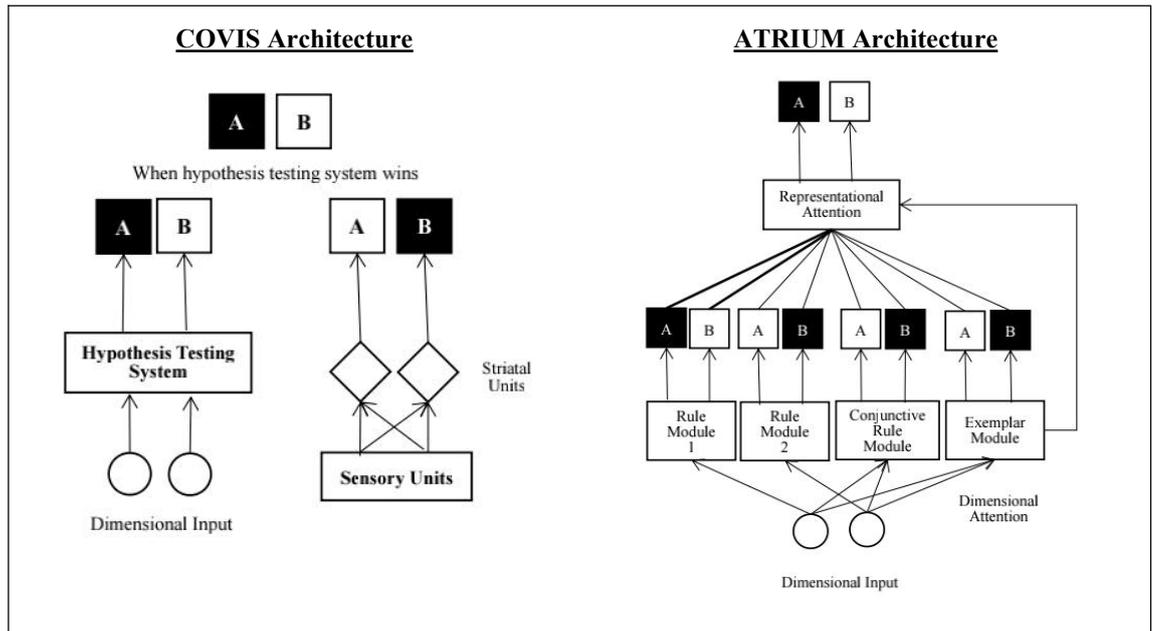


Fig 2-4. (A): Schematic illustration of the architecture of the parallel dual systems framework of COVIS; (B): Schematic illustration of the architecture of the mixture-of-experts framework of ATRIUM (darker lines represent stronger representational attentional bias to the representational module).

2.2.3.3 Interaction Between Multiple Systems

COVIS has correctly predicted many of the empirical dissociations between RB and II category learning that have been reported (Ashby et al., 2011; Helie et al., 2012a; 2012b), including both behavioural and neuroscience data. As a dual systems theory, how does COVIS account for interactions between multiple systems? As mentioned above, COVIS assumes that distinct systems are competitive. The competition between the systems is driven by a separate heuristic that combines the long-term accuracy of each system (known as the “trust” term) with the decisiveness of each system regarding the current stimulus (known as the “confidence” term). This mechanism does not seem to allow back and forth trial-by-trial system switching.

In addition, Ashby and Crossley (2010) proposed that, in a single categorisation task, learning in one system inhibits responses from the other system. Crossley and Ashby (2015)

further revealed that learning in the procedural learning system may occur even while the explicit learning system is in control of a categorisation responses. In the Crossley and Ashby (2015) experiment, a “parse congruent” condition was introduced in which three phases of training used the same II categories but some stimuli were never shown in phases 1 and 2. During phases 1 and 2, stimuli were selected from the two categories in such a way that a 1D rule could achieve optimal accuracy. When provided with immediate feedback during phases 1 and 2, performance in phase 3 of the parse congruent condition was similar to the control congruent condition. However, when provided with delayed feedback during phases 1 and 2, performance in phase 3 of the parse congruent condition was significantly impaired. The result suggests that 1) multiple representations in category learning are possibly processed in parallel; and 2) separate representations or systems inhibit one another to access the output response generation. This lateral inhibition principle is meaningful for task switching, but, unfortunately, it has not been implemented within the COVIS model.

2.2.4 Mixture-of-Experts Architecture

A classic assumption in the early category learning literature was that category representations in a single task are homogeneous and integrated. Hence, much of the effort in traditional categorisation research was focused on understanding the function that maps the input to the output in terms of a unitary process. However, more recent investigations suggest that this homogeneous representation hypothesis may be wrong. In fact, there exist many ill-defined categories in our daily life, just like category structures of task Type III, IV and V in Shepard et al. (1961) in which a set of stimuli belonging to the same category could be composed of multiple subsets (i.e., rule-plus-exception). Erickson and Kruschke (1998), in their empirical studies, showed that people can effectively learn rules with a few exceptions, and generalise from such learning in a rule-like fashion in the test phase. Note that here an explicit rule-based representation and an exemplar-based representation act in parallel, though they are qualitatively different. In other words, the representations people form in a single categorisation task may

involve multiple sub-representations. Several studies have confirmed this hypothesis (Denton et al., 2008; Sewell & Lewandowsky, 2011; Yang & Lewandowsky, 2003; 2004).

ATRIUM (attention to rules and items in a unified model) (Erickson & Kruschke, 1998) is another extension of the attention learning framework that accounts for heterogeneous category representations. Like COVIS, ATRIUM also emphasises that category learning is determined by distinct representational modules: a rule-based representation and an exemplar-based representation. With ATRIUM, Erickson and Kruschke (1998) introduced a mixture-of-experts architecture. Instead of using pure competition between different representational modules, the ATRIUM model incorporates a gating network, in which each category decision is influenced by the activation level of the gate node corresponding to each module (see Fig 2-4B). A gate node is associated with the exemplar nodes within the exemplar network module: the association weights between the exemplar nodes and gate nodes are adjusted up and down in terms of each module's accuracy on a trial-by-trial basis. As the gating network coordinates different representational influences, it is also referred to as *representational attention* (Craig & Lewandowsky, 2012; 2013; Erickson, 2008; Erickson & Kruschke, 1998; Lewandowsky, 2011; Sewell & Lewandowsky, 2011).

ATRIUM was motivated by a unifying mathematical algorithm, whereby the rule network, the exemplar network and the gating between them are all driven by the delta learning rule—gradient descent on the error. One advantage of ATRIUM is that the category decision produced by this model can be stimulus-specific. This principle does facilitate the formation of heterogeneous category representations, though it may be problematic in some cases of large size category structure (e.g., Ashby & Crosley, 2010; Crossley et al., 2017). In addition, Lewandowsky and his colleagues (Craig & Lewandowsky, 2012; Erickson, 2008; Lewandowsky et al., 2012; Sewell & Lewandowsky, 2012) proposed that, as it was found that the capacity of regulating the multiple sub-representations in a heterogeneous category learning task is strongly associated with the higher-level cognitive processes, the representational attention implemented in the gating network may also reflect some properties of cognitive control.

2.2.5 Three Common Principles

The initial proposals of multiple representations of category learning (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Love et al., 2004) share three common principles. First, all of them postulate that RB tasks and II tasks are dissociable. There may be one system that subserves category learning by selectively shifting dimensional attention, and forming a rule-based representation, while another system subserves category learning by using perceptual similarity, and forming an exemplar-based representation (e.g., Ashby & Maddox, 2010; Sloutsky, 2010). Neurologically, formation of these representations is associated with different brain regions. A widely accepted idea is that formation of the rule-based representation is primarily associated with frontal cortex, while formation of the exemplar-based representation is primarily associated with basal ganglia (BG) (e.g., putamen) and the premotor area (e.g., Ashby & Maddox, 2010; Ashby & Valentin, 2016; Rogers et al., 2000).

Second, the rule-based representation is deployed by default in category learning. This idea stems from evidence that participants exhibit more flexible and faster learning in RB tasks than in II tasks (e.g., Ashby, et al., 1998; Kruschke, 1993). In addition, normally functioning adults tend to categorise according to verbal rules and only switch to the similarity-based system when it is clear that no acceptable verbal rule exists. Even for an II categorisation task, for which no verbal rule exists, participants tend to use a verbal rule early in learning but switch to a similarity-based strategy as learning progresses (e.g., Maddox et al, 2004; Markman et al, 2006; Worthy et al, 2009).

Finally, almost all of the initial proposals hold that multiple representations compete, and possibly inhibit one another, to determine the response on a given trial. The competition hypothesis comes from neuroimaging studies which have found an antagonistic relationship between hippocampal and striatal activity (e.g., Dagher et al., 2001; Jenkins et al., 1994; Nomura et al., 2007). Animal lesion studies have also revealed that medial temporal lobe lesions can improve performance in striatal-mediated learning tasks, while striatal lesions can improve performance on medial temporal lobe dependent tasks (e.g., Mitchell & Hall, 1988; O'Keefe & Nadel, 1978).

2.3 Implied Cognitive Control In Category Learning

2.3.1 Automatic and Non-Automatic Control

Most theories of automatic behaviour were developed before the widespread development of multiple representations theories of category learning (e.g., Crossman, 1959; Ericsson et al., 1993; Palmeri, 1997). A widely accepted theoretical framework is that the control system of automatic behaviour is independent of the control system of the non-automatic or deliberative behaviour (e.g., Cohen et al., 1990; Norman & Shallice, 1986; see also Cooper, 2010 as a brief review). A prototypical account is Norman and Shallice's (1986) informal model of non-automatic control, the Supervisory System theory. According to the Supervisory System theory, with more and more practice, automatic behaviour develops and control is transferred from the deliberative system to the automatic control system, so-called Contention Scheduling System. But Supervisory System is able to modulate the operation of Contention Scheduling, and hence well-learned behaviour, when desired or necessary.

Given that, according to COVIS, category representations are formed in multiple systems and the exemplar-based system is independent of cognitive control, how might the theoretical integration of the multiple representations theory and the cognitive control theory be achieved? Recent research on automatization provides an intriguing starting point. Not surprisingly, Ashby and colleagues recently reported results from several categorisation studies that support the view that there should be only a single system that controls automatic behaviour (see some results from Helie (2010) shown in Fig 2-5) (Helie et al., 2010a; 2010b; Soto et al., 2013; Waldschmidt & Ashby, 2011; see also Ashby & Crossley, 2012 for review). In these studies, participants were trained either on RB or II category structures for more than 10,000 trials. Although many behavioural dissociations between RB and II categorisation were observed in the early stage of training, all of them disappeared after 8,000 trials. For example, Helie et al. (2010a) reported that statistically significant differences on accuracy and RTs between RB learners and II learners disappeared (Fig 2-5B and C), though the acquisition of II categories required a longer period of practice in the beginning. Switching the response locations after automaticity has been developed impaired both RB and II tasks (on both RTs and accuracy levels) (see Fig 2-5D), and

there was no recovery from this interference over a long period of time on practice. However, adding a secondary task did not significantly interfere with either performance measure on either type of categorisation task (see Fig 2-5E).

The automatic control system is different from both the explicit learning system and the procedural learning system. Neuroimaging studies have also revealed that in the early stage of category learning, the RB task is more correlated to activity within the PFC and head of caudate nucleus, whereas the II task is more correlated to activity within the putamen (e.g., Milton & Pothos, 2011). However, after extended training, only cortical (premotor cortex) activation is confirmed to be correlated with automatic categorisation performance (e.g., Helie et al., 2010; Soto et al., 2013; Waldschmidt & Ashby, 2011). Perhaps more interesting evidence for this position is that the category representations used in automatic categorisation are similar to the representations formed in category learning. Roeder and Ashby (2016) examined 27 participants who completed 12,300 trials of learning on either RB or II categories. During their experiment, each participant mainly practiced a primary category structure, but every third session they switched to a secondary structure that used the same stimuli and responses. In the switching session, half of the stimuli retained their same response on the primary and secondary categories (congruent) and the rest switched responses (incongruent). Participants' performance on the primary categories met the standard criteria of automaticity by the end of practice. But, they found that, in the RB condition, differences between the accuracy and response time (RT) on congruent and incongruent stimuli were not statistically significant, whereas in the II condition, mean accuracy was significantly higher and RT was lower for congruent than for incongruent stimuli. Roeder and Ashby thus concluded that RB and II categories were automatized differently, the RB task was automatized by abstracting a rule and the II task was automatized through S-R mappings. Some studies have reported that in addition to response-sensitive neurons, rule-sensitive neurons are also found in premotor cortex (e.g., Muhammad et al., 2006; Wallis & Miller, 2003). The results suggest that automatization of the RB task can be seen as transfer of control from explicit learning circuits (i.e., PFC, ACC and caudate nucleus) to rule-sensitive neurons of premotor cortex, while automatization of II tasks consists of transfer of control from the putamen to response-sensitive neurons in the premotor cortex.

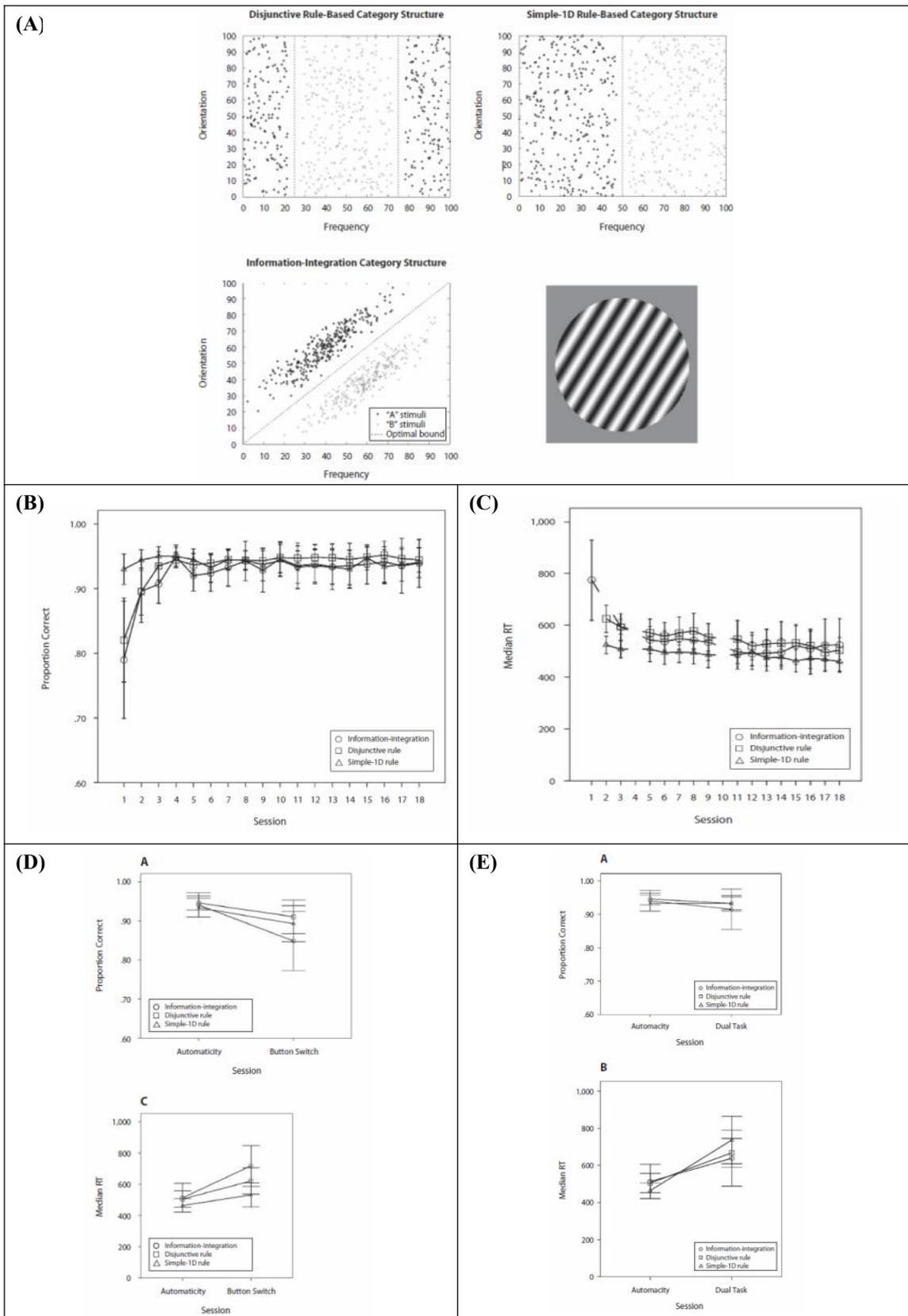


Fig 2-5. Results from Helie et al. (2010a). (A) shows category structures and example stimulus used in the experiments. (B) and (C) show proportion correct and mean median correct reaction time (RT) per training session, respectively. (D) shows proportion correct and mean median correct RT for automatic categorisation performance and during the button switch session. (E) shows proportion correct and mean median correct RT for automatic categorisation performance and during the dual task interference session. (The error bars are 95% confidence intervals.)

Indeed, the single automatic control system and the multiple representations theory of category learning are not necessarily in conflict. RB tasks and II tasks can be controlled by a single automatic control system after extensive training with the distinct representations to be stored. To move toward an integrated account, it is necessary to rethink the role(s) of non-automatic control, or known as cognitive control, in the multiple representations of category learning. Although it has been argued that the II category learning might occur in the absence of cognitive control in some cases, this does not mean that cognitive control is not involved in the processing of competition between multiple representations.

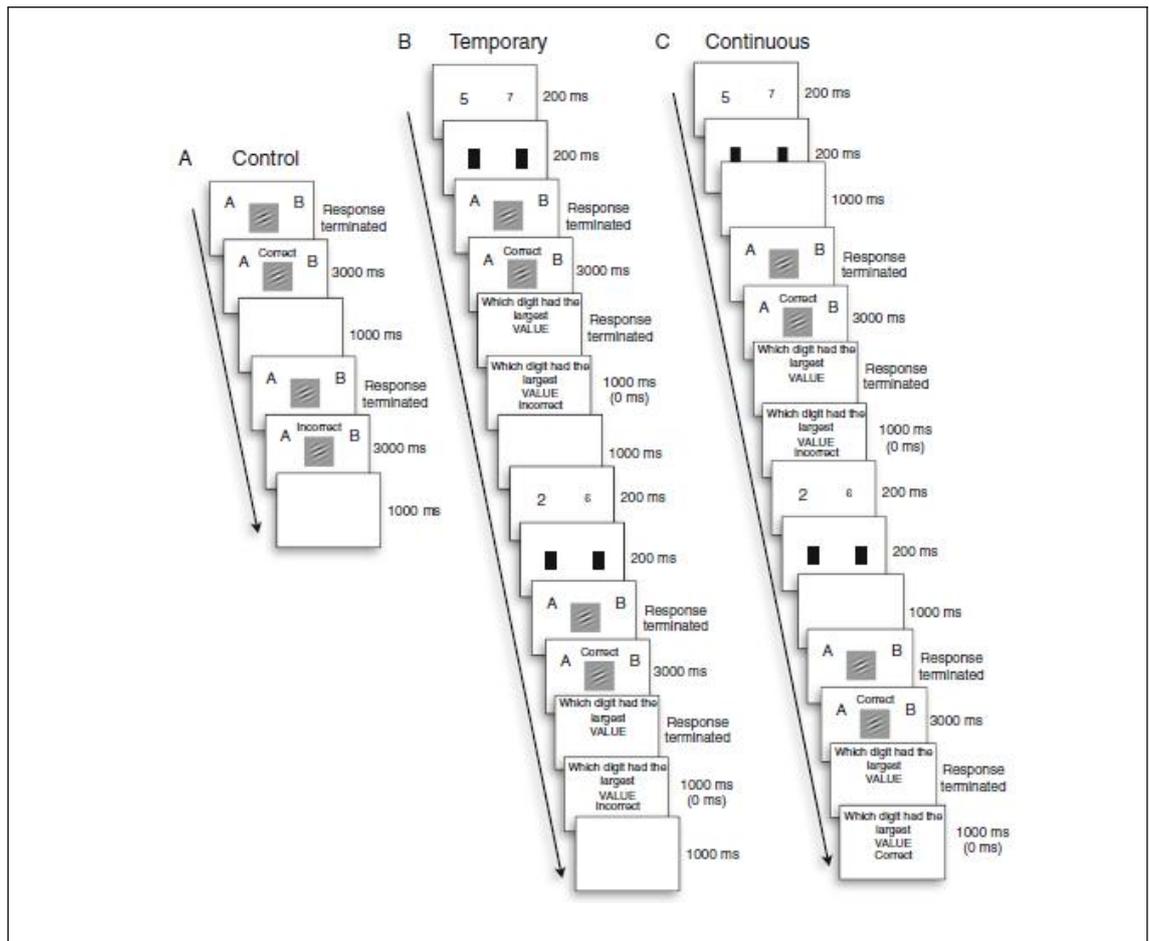
2.3.2 Information-Integration Category Learning and Cognitive Control

Once the multiple representations hypothesis is accepted in the domain of categorisation, the next question, naturally, to ask is how the different representations interact. According to COVIS, prefrontally-mediated cognitive control is important for RB tasks, but II tasks could be independent of cognitive control (e.g., Ashby et al., 1998; Ashby & Maddox, 2010; Ashby & Valentin, 2016). The traditional research has established that the cognitive control-dependent and -independent systems compete to control action selection according to task demand. However, more recent research has provided considerable evidence against the cognitive control independent hypothesis. Much recent research has shown that the independent learning hypothesis may be wrong (e.g., Paul & Ashby, 2013).

Although the exemplar-based system seems not to rely on cognitive control to make decisions or process feedback, it may not be the case that the II category learning operates totally independently of cognitive control. The first branch of research on representational interaction is the investigation of system transitions. Research on system transition derives from the issue of how control of behaviour transfers from rule-based system to exemplar-based system during II category learning. Given that the rule-based system is deployed as default in the beginning of

category learning, how could II learning performance be optimised by overcoming the inhibitory control from the rule-based system and a default bias to achieve the asymptotic performance?

The RB tasks are mediated by cognitive control, whereas the exemplar-based system operates using relatively basic cognitive processes (Helie et al., 2010; Maddox et al., 2004; Zeithamova & Maddox, 2006; 2007). Thus, once the exemplar-based system is in control, it would have no need for cognitive control processes. Even so, it is possible that cognitive control plays a role in facilitating the transition away from the rule-based system, because in the literature it has been confirmed that there is a tendency to use simple rule-based strategies during initial learning. For example, in II category learning research, the tendency to use suboptimal rule-based strategies has often been observed, though these strategies produce lower accuracies (e.g., Ashby et al., 2003; Maddox et al., 2004). Interference manipulations such as delayed feedback and extreme levels of category overlap can also lead to an increase in the use of suboptimal strategies (e.g., Ashby & O'Brien, 2007; Ell & Ashby, 2006; Smith et al., 2014). On the other hand, much recent research has revealed that cognitive control may be correlated with the performance of II category learning. While, unlike rule selection and maintenance that are primarily associated with prefrontal cortex, II category learning is mainly associated with the basal ganglia, it may not be the case that II tasks are unaffected by cognitive control. Given that the rule-based system is dominant, it is reasonable to believe that cognitive control may be important for transitioning from the rule-based system and engaging the exemplar-based system.



D. Model-Based Strategy Identification

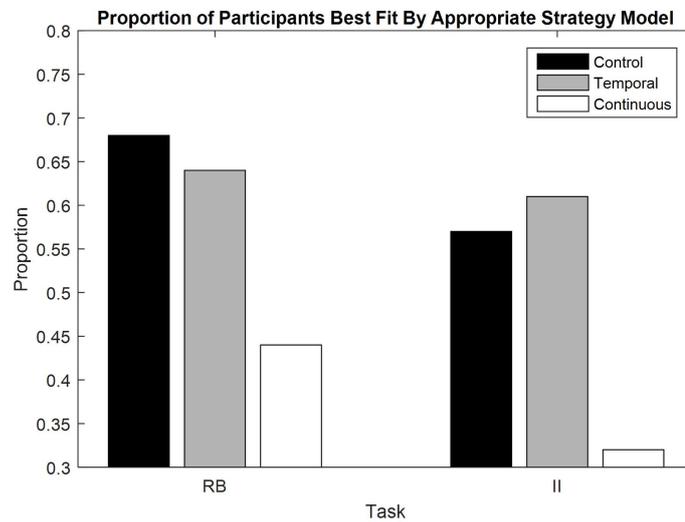


Fig 2-6. Top: Task design for the first 10 trials, with timing for trials 11–320 given in parentheses, used in Miles et al. (2014). A Control condition. B Temporary condition. C Continuous condition. Note that the temporary and continuous conditions are the same, except that the intertrial interval was before the presentation of the digits in the temporary condition and after the presentation of the digits in the continuous condition. Bottom: D Proportion of participants best fitted by RB and II strategy models.

In fact, a series of recent studies have implied that cognitive control is important for performing II category learning. Participants with frontal lobe damage (Schnyer et al., 2009), older adults (Maddox et al., 2010b), young children (Huang-Pollock et al., 2011), and sleep-deprived adults (Maddox et al., 2009) were trained on II categories. Note that, in all cases it was expected that these groups of participants would show normal performance, relative to controls, though they all have decreased cognitive control. But, surprisingly, decreases were found in these groups, suggesting that cognitive control might be involved in II category learning. Mathematical model-based analysis of the participants' categorisation strategies, in addition, illustrated that these groups tended to use a suboptimal RB strategy to solve the II task, suggesting difficulty making the transition away from the dominant rule-based system. Moreover, it was found that the tendency to use the exemplar-based system to perform the II category learning task was often positively related with the capacity of some cognitive control processes (e.g., Maddox et al., 2010; Schnyer et al., 2009).

Previous research has also shown that concurrent secondary tasks that tax some cognitive control processes interfere more with RB categorisation than with II category learning tasks (e.g., Miles & Minda, 2011; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006; 2007). However, in the more recent literature, Miles et al (2014) found that continuously taxing cognitive control using concurrent numerical Stroop task (for a detailed review of the Stroop task see MacLeod, 1991) has a different effect on the II category learning task than does temporary secondary task interference (see Fig 2-6). In traditional concurrent secondary task studies, the intertrial interval was given before presentation of the digits. As can be seen in Fig 2-6, instead of placing a blank screen before the presentation of each secondary task stimulus (see 2-6B), Miles et al. (2014) placed each blank screen after the presentation of each secondary task stimulus (see 2-6C). This change caused a significant decrease of using an II strategy in II category learning (see 2-6D). This is consistent with the idea that when EFs were never fully available – the II categorisation task was less likely to be learned using the appropriate strategy, suggesting that continuously taxing cognitive control processes interferes with the transitioning from a rule-based representation system to the exemplar-based representation system. Note that this is by no means to say the process of transitioning from a rule-based representation to an exemplar-based

representation is the same, but it is plausible to suggest that cognitive control has some role to play in the interaction between different representational systems.

To summarise, taken together, recent research on category learning has confirmed that there are multiple representations which are formed dissociatively in the brain and can be computationally represented by different formats. However, how these competing representations interact remains debatable. Results from much recent research implied that cognitive control plays an important role in the interaction between multiple representation systems. However, because of the limitations of the traditional paradigm, many of these results cannot allow theoretical development towards an integrated computational account of cognitive control in category learning. In the next section, a new paradigm of category learning that allows task switching in categorisation will be introduced.

2.3.3 Heterogeneous Categorisation and Task Switching

2.3.3.1 Heterogeneous Categorisation and the Task Partitioning Paradigm

Heterogeneous categorisation tasks are those in which participants need to learn to classify exemplars into multiple substructures with incompatible boundaries. Aha and Goldstone (1992) tested 40 undergraduate participants by using the category structure shown in Fig 2-7A. In this figure, horizontal and vertical axes denote (increasingly right) line position and (decreasing) square size dimensions respectively. Items labelled A and B are 12 training items belonging to different categories, while those labelled X, Y, W and Z are critical test items. All participants completed training with 4 error-free classifications of the complete set of training items. In the test phase, participants were found to apply two separable rules to classify stimuli (see Fig 2-7B).

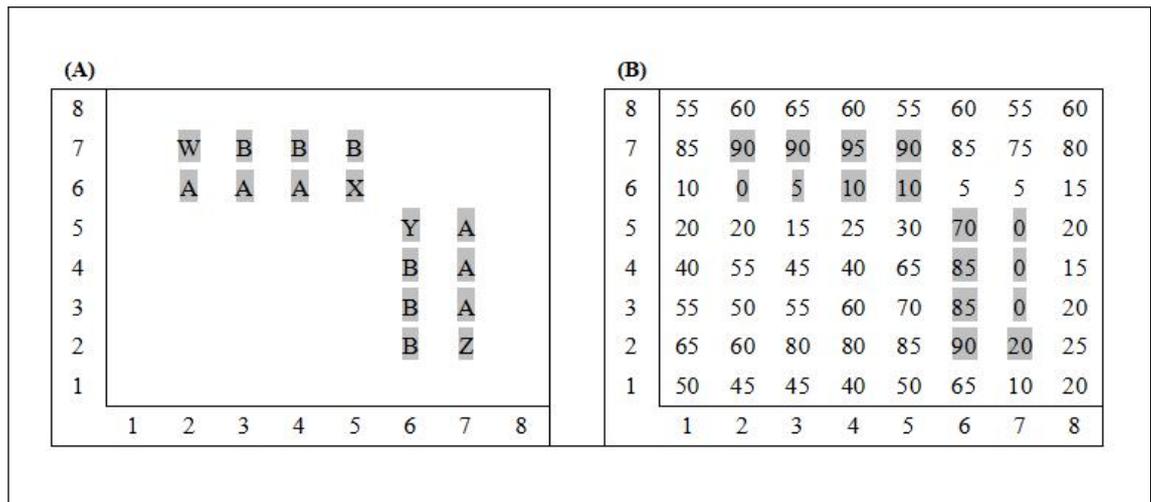


Fig 2-7. (A). Training sets (shown as A and B) and critical test items (W, X, Y, and Z) used in Aha and Goldstone (1992); (B). The proportion of category B responses in the transfer phase. The horizontal axis denotes line positions and vertical axis denotes square size.

Following Aha and Goldstone (1992), Lewandowsky and his colleagues (2003; 2004) developed the task partitioning paradigm, also known as knowledge partitioning, in category learning. In the task partitioning paradigm, stimuli are presented with a normatively irrelevant contextual cue (e.g., background colour) to signal the presence of a local regularity in the partitioned task environment. It is held that when task partitioning occurs, participants simplify the complex task by decomposing it into separate local solutions. Yang and Lewandowsky (2003; 2004) trained their participants with the heterogeneous category structure with separate contextual cues. They confirmed that their participants can perfectly learn the structure. Other investigations have further confirmed that the local solutions in a heterogeneous category structure can be modulated either by contextual cues or stimulus components (e.g., Erickson, 2008; George & Kruschke, 2012; Sewell & Lewandowsky, 2012).

Most previous task partitioning studies focused on category structures that only involve some subsets of incompatible verbal rules. This is somewhat analogical to the switching version of the Stroop task (see Allport & Wiley, 2000; Gilbert & Shallice, 2002). In the Stroop task (Stroop, 1935), stimuli are colour words presented in coloured ink (e.g., the word RED presented in blue ink), and participants must name the ink colour. Responses are slower when the colour name and ink colour are different (incongruent trials) than when they are the same (congruent trials). For a full review of effects, see MacLeod (1991). In the switching version, participants are required to name the colour of the ink on some trials but to name the colour word on other

trials (i.e., to switch between colour naming and word reading). Interference (as measured by slowed reaction times) is greater when switching from colour naming (the harder task) to word reading (the easier task) than when switching the other way. Such effects are robust in some analogical paradigms that require switching between incompatible rules (Kiessel et al., 2010; Allport et al., 1994). To this end, since task partitioning requires switching between category representations, it may not be surprising that task partitioning could cause switch effects. However, a question to ask is if the task partitioning paradigm can be applied to the heterogeneous category structure consisting of RB and II substructures?

2.3.3.2 Hybrid Category Learning

Ashby and Crossley (2010) introduced a hybrid category learning task in which a 1D RB strategy is required on some trials and an II strategy is required on other trials (see Fig 2-8A). Participants received two sessions (600 trial for each) of training. For comparison purposes, in the first session, three groups of participants were trained in RB, II, and hybrid category structure. Ashby and Crossley (2010) reported that, no matter in which condition, participants have difficulty learning the hybrid strategies. Instead, the strategy model analysis showed that the participants tended to use the 1D rule-based strategy for all trials (see Fig 2-8B). This suboptimality was explained by suggesting that the failure of task switching may be due to lateral inhibition between separate category learning systems.

Theoretically, it is not a problem to have lateral inhibition between separate representation systems. In recent literature, considerable evidence of lateral inhibition effects on the interaction between separate representation systems has been reported. For example, neuroimaging evidence has found an antagonistic relationship between neural substrates for declarative and nondeclarative memory systems (Poldrack et al., 2001; Schroeder et al., 2001). However, it is problematic that in Ashby and Crossley's (2010) empirical approach they did not use the task partitioning paradigm. The category representations in the hybrid structure are not heterogeneous, but homogeneous. At least, category A can be easily achieved by using either an RB strategy or an II strategy, because the boundaries are compatible. Participants do not need to learn to switch between rules, but to excite one system and inhibit the other to achieve an integrated task set (e.g., Ashby & Crossley, 2010; Ell & Ashby, 2006).

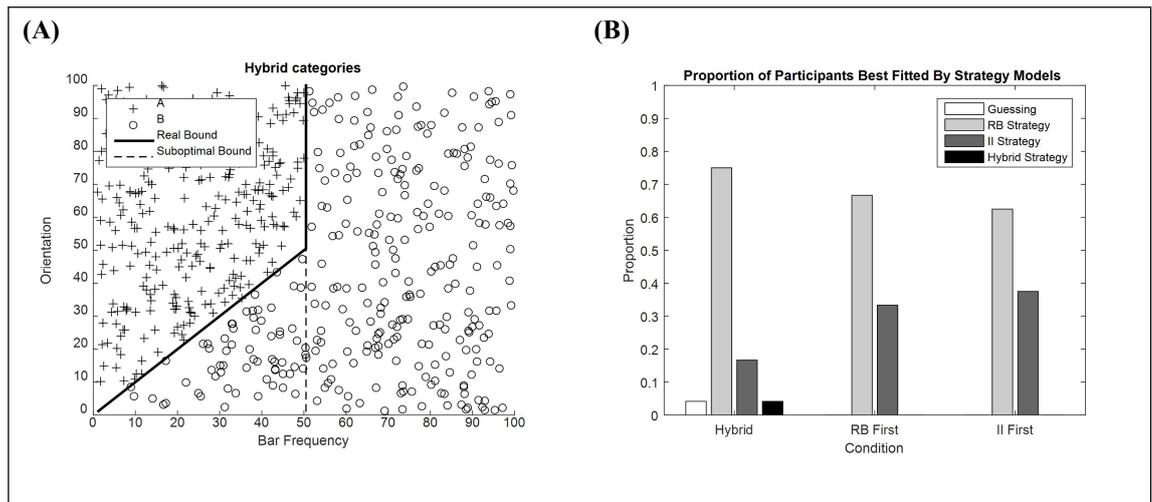


Fig 2-8. (A): Schematic illustration of Hybrid category space used in Ashby and Crossley (2010) Experiment. (B): Proportion of participants best fitted by each strategy model in session 2.

2.3.3.3 Task Switching Effects

Intuitively, switching between RB and II tasks may be an essential capability in human cognition. As we discussed earlier, an ideal II category learning process can be seen as a gradual transition from a rule-based representation system to an exemplar-based representation system. However, the reason why we learn is to apply the knowledge we have learned. To deal with the changing environment, we may need to flexibly switch between tasks in response to environmental requirements.

Erickson (2008) introduced the task partitioning paradigm into category learning with a heterogeneous category structure consisting of an RB substructure and an II substructure. In Erickson's (2008) experiment, participants were required to go through three training phases. In the first two phases, participants, respectively, learned an RB task and an II task. Afterwards, participants learned to switch between those tasks on an intermixed, trial-by-trial basis. Erickson (2008) reported that when more switching cues (i.e., adding background colour) were provided and the II task had different responses than the RB task (i.e., A and B versus C and D), a significant proportion (more than 40%) of participants showed the capacity of trial-by-trial switching between RB and II substructures. More importantly, the RT switch costs of successful learners (defined by strategy model fitting) were greater than of non-learners (see Fig 2-9B).

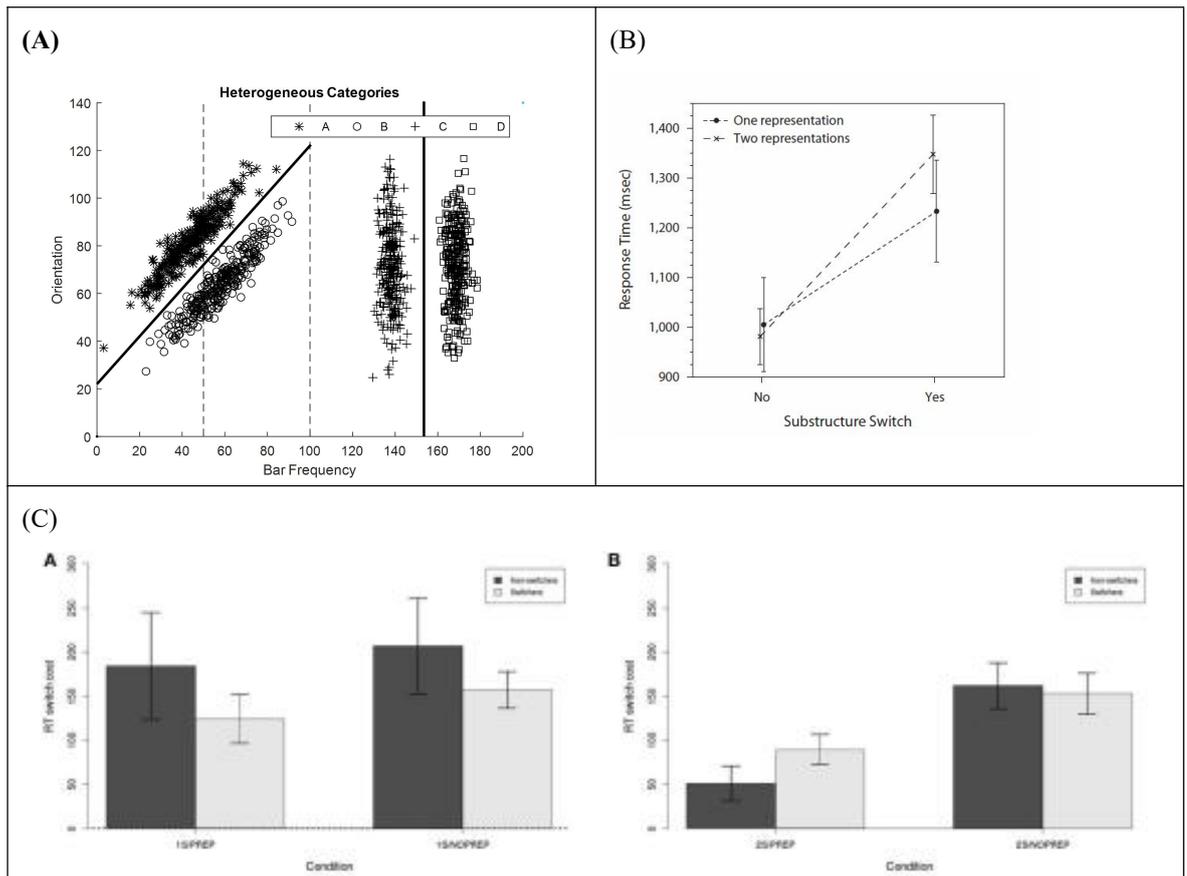


Fig 2-9. (A): Schematic illustration of a heterogeneous category structure. (B): RT costs for different trial types reported by Erickson (2008); (C): Switch costs in different conditions reported by Helie (2017).

Crossley et al. (2017) extended Erickson’s paradigm by comparing the RB and II switching conditions with an additional condition in which participants learned to switch between two RB tasks (i.e., 1D rule versus XOR). They also introduced a button switch interference in the final 100-trial block of trial-by-trial switching and found that the interference effects were correlated with the task difficulties. Accuracy recovered in the two RB tasks but did not recover in the II task. Although, half of the participants in the task switching condition performed well, more participants abandoned the optimal strategy for the II task than the XOR task during the trial-by-trial switching, and the switch costs were significantly greater than switching between two RB tasks. Following Crossley et al. (2017), Helie (2017) further introduced a factor, preparation time, into task switching in heterogeneous category learning. Interestingly, the results showed that with practice, the effects of preparation time on facilitating task switching in heterogeneous category learning become significant (see Fig 2-9C).

Taken together, the research of task switching shows that trial-by-trial switching between RB and II tasks seems to be more difficult than RB task switching, but it is possible. This

suggests that cognitive control should also play an important role in the interaction between multiple representation systems, though it may be very effortful. This seems a little inconsistent with the past assumption that switching between representations on a trial-by-trial basis is a common and routine occurrence (Ashby et al., 1998; Erickson & Kruschke, 1998). Extensive training is required.

2.4 Concluding Comments

Almost all of the empirical and computational approaches of perceptual categorisation have already confirmed that there are multiple internal representations involved during category learning. Based upon the multiple representations theory, recent empirical evidence indicates the need for a control mechanism to participate in categorisation to mediate the organisation of multiple internal representations. Recent research on categorisation automaticity, II category learning and heterogeneous category learning, has shown that cognitive control may play an important role in the control of the organisation of multiple representations. In particular, the heterogeneous category learning research implying task switching does allow the theoretical development.

However, there is a gap between the empirical and computational approaches. That is, while most recent computational approaches address the role of attention in category learning, they do not consider the role of attentional control. This may be because the primary focus of previous computational modelling of category learning was originally on establishing that humans are able to learn and use multiple representations within a single categorisation task, and not on building a model to precisely predict the control mechanisms mediating the organisation of multiple internal representations. Given that the internal representation developed in category learning is associated with the distribution of dimensional attention, and given the competition between the multiple representations, it is adaptive to have a control mechanism to carry out this complex task.

How do we interpret the role of cognitive control in categorisation? The Supervisory System theory of attentional control (Norman & Shallice, 1986) may provide a solution, because,

in line with the evidence from the research of categorisation automaticity, the Supervisory System theory addresses that control of automatic behaviour and non-automatic behaviour as being mediated by distinct systems. Automatic behaviour is controlled by a system called Contention Scheduling, whereas the non-automatic behaviour is controlled by the Supervisory System. In addition, it is argued that the Supervisory System can modulate Contention Scheduling when required. Supervisory System is not always in control, with more and more practice, the behaviour is gradually controlled by the Contention Scheduling with less and less mediation from the Supervisory System. This account has been successfully applied to explain multitasking (Gilbert & Shallice, 2002) and the organisation of daily behaviour (e.g., beverage preparation) (e.g., Cooper et al., 2014; Cooper & Shallice, 2000). Moreover, Erickson (2008) argued that the Supervisory System account, at a conceptual level, does share some properties with the control of multiple representations in category learning. Therefore, it is proposed in this thesis to combine the attentional control theory with category learning models to account for the role of cognitive control in categorisation.

It is necessary to revisit those existing computational accounts before developing an integrated account of cognitive control over the organisation of multiple internal representations. A computational exploration is therefore in order. In the following four chapters, reimplementations of the models described above will be described, and their advantages and limitations evaluated. Based on the operational mechanisms of these models, with the empirical findings mentioned above, a new modelling scheme will then be proposed.

Chapter 3.

Attention to Categorisation: The Perspective from Exemplar Theory

3.1 Introduction

Exemplar theory assumes that a category response to a stimulus input is determined by the stimulus' similarity to the learned exemplars of that category, relative to its similarity to exemplars of the other categories. The similarity is represented as a monotonically decreasing function of distance in a multidimensional psychological space. The most dominant account that incorporates the principles of attention allocation and exemplar similarity is the GCM model (Generalised Context Model, Nosofsky, 1986). The key assumption in GCM is that people gradually learn to assign attention to different stimulus components so as to achieve optimal performance. For example, let $[x_{j1}, x_{j2} \dots x_{jn}]$ denote the values on n dimensions of the current stimulus j , let $[y_{i1}, y_{i2} \dots y_{in}]$ denote values on n dimensions of a learned exemplar i of category A . Then, the similarity $S_{ij}(A)$ is given by

$$S_{ij}(A) = \exp(-c(\sum_n a_n |x_{jn} - y_{in}|^r)^{1/r}) \quad (3.1)$$

where $c (> 0)$ is the tuning of the receptive field. If c is large, similarity decreases rapidly with increasing distance, whereas, if c is small, similarity decreases slowly with increasing distance. a_n is the attention strength given to dimension n . r is a constant determining the psychological-distance metric. When a separable psychological distance is assumed, r is set to 1, resulting in a city-block metric. However, when an integral psychological distance is assumed, r is set to 2, resulting in a Euclidean metric. Nosofsky (1986) argued that increasing the attention strength on one dimension has the effect of stretching that dimension, so that the distances along that dimension will have a larger influence on the similarity. This attentional flexibility is critical

for stretching dimensions that are relevant for responding, and shrinking dimensions that are irrelevant.

The selective attention mechanism in GCM is empirically successful (see Nosofsky, 2011 for review). Kruschke (1992), thus, combined the selective attention mechanism and a feed-forward back-propagation approach to produce the first network model, ALCOVE (Attention Learning COVERing mapping), of category learning. ALCOVE has accommodated a variety of empirical phenomena in homogeneous category learning with great quantitative precision (e.g., Kruschke, 1992; 1996a; 1996b; Lewandowsky, 1995; Nosofsky et al., 1994a). According to Kruschke (1992), the key role for attention is how much each stimulus dimension is used in the calculation of similarity. In some cases, categorisation performance could be explained by the dimensional focus of attention, where attention strength reflects the relevance of each component of a stimulus. For example, much evidence shows that categories with fewer relevant dimensions can be learned faster than those with more relevant dimensions (Nosofsky et al., 1994a; Kruschke, 1993), while attention shifts to relevant cues are much easier (i.e., more reliable) than to irrelevant cues (Kruschke, 1996a). However, the weakness of ALCOVE has been confirmed in cases where categories are composed by multiple heterogeneous representations. For example, when the category involves a set of stimuli, most of which are categorised according to a rule but some of which are exceptions, human participants can learn to shift between strategies, and this leads to extensive behavioural shifts in the transfer phase. Some previous research has shown that ALCOVE fails to account for these cases (e.g., Denton et al., 2008; Erickson & Kruschke, 1998; Goerge & Kruschke, 2012; Yang & Lewandowsky, 2003; 2004).

Although the ALCOVE model has shortcomings in heterogeneous categorisation tasks, it is the originator of many multiple representations network models. The operating mechanisms of three well-known feed-forward network models—SUSTAIN (Love et al., 2004), COVIS (Ashby et al., 1998), and ATRIUM (Erickson & Kruschke, 1998)—are more or less derived from it. In particular, as discussed in Chapter 2, both ATRIUM and SUSTAIN retain the selective attention mechanism. ALCOVE has also been extended in many ways. For example, it was extended by Kruschke (1996) to produce the AMBRY model (an ambry is, as Kruschke, 1996, notes, a special kind of alcove), which is applied to account for effects on response reversal in category learning. In contrast, previous computational modelling research has long treated ALCOVE as a representative of the single representation/system theory (e.g., Kalish et al., 2017; Newell et al., 2010; Nosofsky & Johansen, 2000; Nosofsky & Kruschke, 2002; Nosofsky et al., 2005), because ALCOVE performs well in quantitatively fitting human data in homogeneous category learning tasks. Therefore, it remains valuable to reimplement ALCOVE so as to have fuller insight into its strengths and weaknesses.

3.2 Reimplementing ALCOVE

3.2.1 Model Description

The illustrative structure of ALCOVE is simple (see Fig 3-1). In ALCOVE, each training exemplar is associated with category responses via learned association weights. During training, ALCOVE learns to adjust the association weights. ALCOVE also learns how much attention to pay to each dimension, which results in the stretching of highly diagnostic dimensions and the shrinking of others that turn out to be less relevant. Dimensional attention strengths and association weights are adjusted via gradient descent on error.

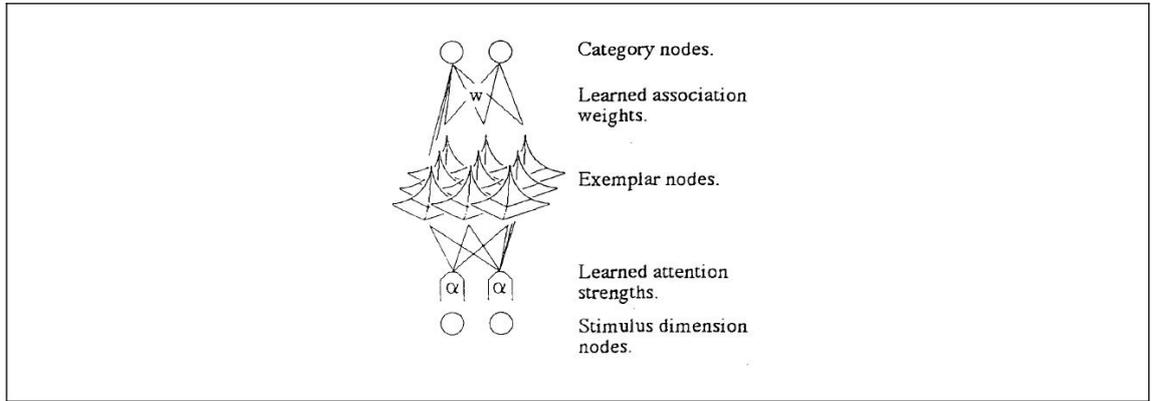


Fig 3-1. Illustrative structure of ALCOVE model.

Exemplars are activated by attention-modulated similarity to the stimulus as described above (Eq 3.1). The activation of each category response is calculated as a weighted sum of all the exemplar activations:

$$a_k^{out} = \sum_j w_{j,k} a_j^{ex} \quad (3.2)$$

where $w_{j,k}$ is the association weight from exemplar node j to category response k . To compare with human data, the activation of category response k needs to be converted to a choice probability. Often, the probability of choosing category k is calculated as follows:

$$P(K) = \frac{\exp(\phi \times a_k^{out})}{\sum_k \exp(\phi \times a_k^{out})} \quad (3.3)$$

where ϕ is a scaling constant, which can be thought of as representing the level of ‘decisiveness’ in the system (see also Love et al., 2004). It may also reflect the strength of lateral inhibition effects between responses.

As mentioned above, learning in ALCOVE is achieved by gradient descent on error. The error is defined as:

$$E = \frac{1}{2} \sum_k (t_k - a_k^{out})^2 \quad (3.4)$$

where t_k is called the ‘humble teacher value’. The teacher values are defined as $t_k = +1$ if the response k is correct, and $t_k = -1$ if the response k is incorrect. Following Rumelhart et al. (1986), association weights and dimensional attention are adjusted proportionally to the error gradient. Evaluating the derivatives leads to the following delta rules:

$$\Delta w_{j,k} = \eta^w (t_k - a_k^{out}) a_j^{ex} \quad (3.5)$$

$$\Delta a_m = -\eta^a \sum_j \left(\sum_k (t_k - a_k^{out}) w_{j,k} \right) a_j^{ex} c |x_{j,m} - y_m| \quad (3.6)$$

The rates at which association weights and dimensional attention strengths are learned are determined by the association learning rate η^w and attention learning rate η^a , respectively.

3.2.2 Example 1: Applying ALCOVE to the Shepard Six Task Types

The data from Shepard et al. (1961) is challenging for computational modelling of category learning. A primary concern of Shepard et al. (1961) was to determine the relative difficulty of learning the six task types. As being defined by a simple one-dimensional rule, task Type I is easiest. However, task Type II is defined by the exclusive-or (XOR) rule, which requires attention to two dimensions, so it is more difficult. Task Types III-VI require attention to all three dimensions, but the dimensions are not equally informative in every type. Task Type VI may require equal attention to all dimensions, while for the other task types, six exemplars can be correctly classified by considering only one of three dimensions, with attention to the other two dimensions for the remaining two exemplars (see Fig 2-1). Therefore, task Type VI should be the most difficult, and task Types III, IV and V should be easier than Type VI, but more difficult than Type II (see also Nosofsky et al., 1994a).

3.2.2.1 Method

Kruschke (1992) applied ALCOVE to mimic learning of the six task types and verified that ALCOVE is able to predict the relative difficulty of each task types. Here, ALCOVE was reimplemented by following the implementation described in Kruschke (1992), except that training of ALCOVE here only consisted of 16 epochs, which is consistent with the number of

training blocks in the behavioural experiment. (Kruschke's (1992) implementation involved 50 epochs.) In this simulation, following Kruschke (1992), the decisiveness constant ϕ was set to 2.00, the specificity constant c was 6.50, and the associative learning constant η^w was 0.03. Two values of η^a (0.0000 and 0.0033), corresponding to no attention learning and moderate attention learning, were considered. The model was coded in MATLAB.

3.2.2.2 Results and Discussion

Fig 3-2 (A and B) show learning curves generated by the reimplementaion of ALCOVE when there is no attention learning (panel A) and when there is moderate attention learning (panel B). The data represent the probability of choosing the correct category, averaged across 16 trials (each exemplar presented twice) in each training block. In the no attention learning ($\eta^a = 0.00$) condition, as can be seen from Fig 3-2A, learning of task Type II is too slow, while, in the attention learning ($\eta^a = 0.0033$) condition, learning task Type II it second fastest.

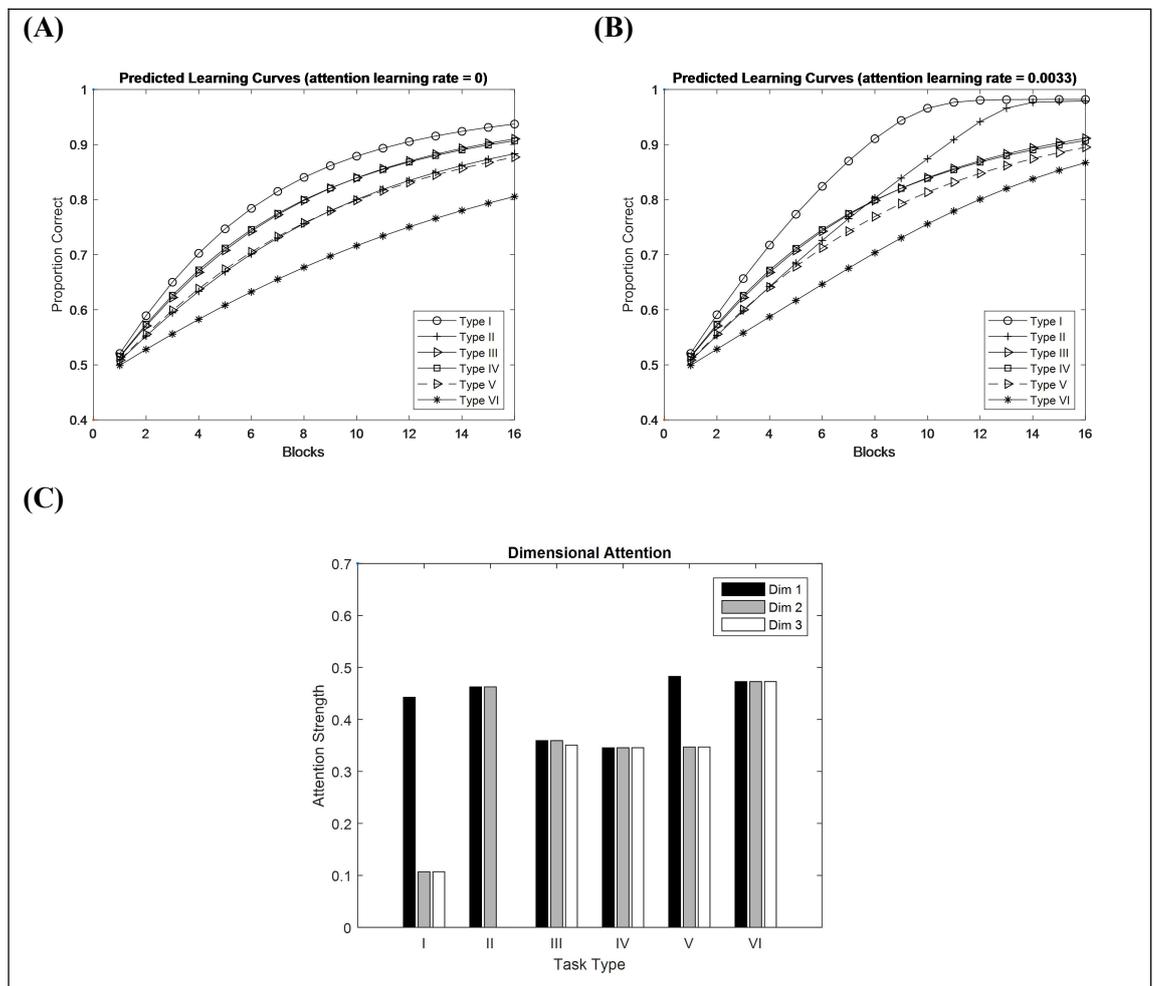


Fig 3-2. (A). Results of applying ALCOVE to Shepard et al. (1961) six task types without attention learning; (B). Results of applying ALCOVE to Shepard et al. (1961) six task types with attention learning; (C). Learned distribution of attention strengths.

Fig 3-2C shows the distribution of learned dimensional attention strengths on each task type. As can be seen from the figure, the distribution of attention strengths obviously varies according to relative difficulty of task types. In particular, in task Type I, attention is mainly assigned to dimension 1, in task Type II, attention is assigned to dimensions 1 and 2, and in task Type VI, all three dimensions. The unequal dimensional attention strengths are not obvious on task Type III and IV, because there are two or more possibilities for solving these task types. But, as there is only one possibility for distribution of attention strength on task Type V, it is obvious that dimensional attention strengths are unequal.

In sum, this simulation has successfully verified that an attention learning mechanism enables a qualitative fit to the classic homogeneous category learning behaviour. Many subsequent studies have further confirmed the power of the selective attention mechanism (e.g., Love et al., 2004; Nosofsky et al., 1994a).

3.2.3 Example 2: Modelling Rule-Based Extrapolation

As mentioned in Chapter 2, there have always been two solutions to learning task Types III, IV and V. The homogeneous representation perspective, like GCM and ALCOVE, suggests that like task Type VI, people learn task Types III-V by remembering every exemplar, but attention strength to each dimension varies. But, the heterogeneous representation perspective suggests that people learn these tasks through the rule-plus-exception strategy. In other words, most items share the membership determined by a simple rule and few items are exceptions (e.g., Kruschke & Erickson, 1994; Love et al., 2004; Nosofsky et al., 1994b).

One influential study of heterogeneous representation is Erickson and Kruschke's (1998) study of rule-based extrapolation. In their experiment 1, Erickson and Kruschke used rectangles varying in height (the primary dimension) and position of an internal segment vertical line (the secondary dimension) (see Fig 3-3B). The Rule-plus-exception category structure used in the experiment is shown in Fig 3-3A. As can be seen in Fig 3-3A, ten of twelve training items could be categorised in terms of the boundary defined by height equal to 4.5 (rule-based exemplars), whereas, two exception items are given which contains open shapes in the figure. Each participant went through 29 blocks of training. Each training block contained 14 trials in which each rule-based exemplar was given once and each exception exemplar was presented twice. After the training trials concluded, participants were tested by letting them assign category membership to transfer stimuli. The exemplars in each corner of the 10 x 10 psychological space were taken as the critical items to distinguish between prediction strategies.

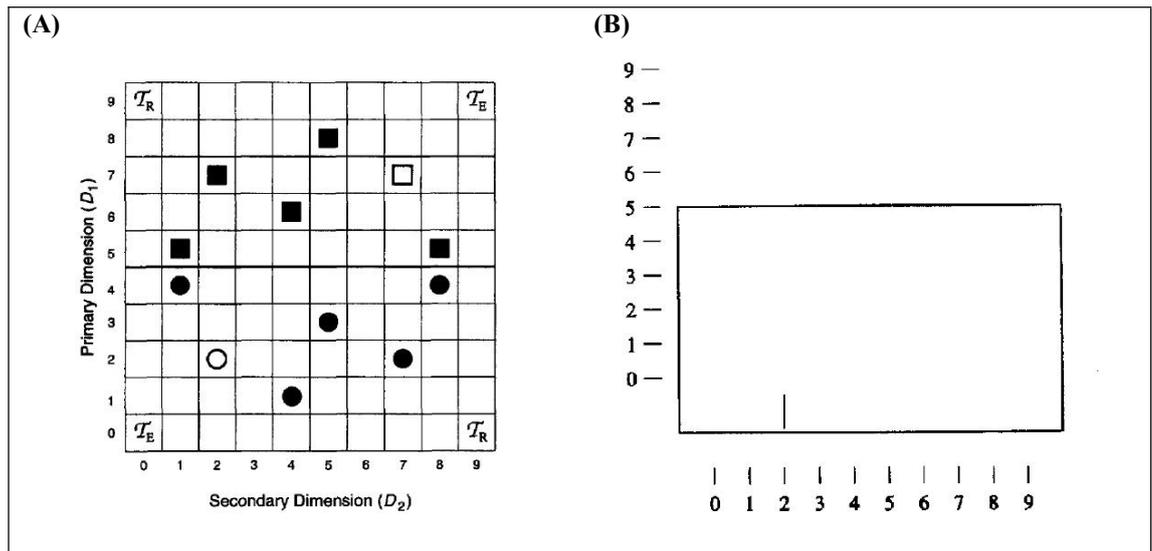


Fig 3-3. Category structure (A) and an example stimulus (B) used in Erickson and Kruschke (1998) Experiment 1. The cells containing filled shapes were rule training instances. Filled squares belong to one rule and filled circles belong to another. The two cells containing open shapes were exception training instances. The cells labelled T_R and T_E indicate the transfer stimuli used to distinguish between predictions from rule-based and exemplar-based modules.

The primary concern of Erickson and Kruschke's (1998) Experiment 1 was to explore the pattern of how people form category representations in the rule-plus-exception task. The results revealed that there exist different types of representations, rule- and exemplar-based. The rule-based representation is demonstrated by examining participants' extrapolation. Erickson and Kruschke (1998) found that, in the test phase, people generalise category knowledge by extrapolating in a rule-like fashion, even when they are presented with a novel stimulus that is most similar to the training exception (see Fig 3-4 right bottom). The exemplar-based representation is inferred by examination of responses to the training exceptions. For example, participants tended to give more exception responses when presented with stimuli that matched the exception on the primary dimension than when presented with stimuli that matched on the secondary dimension. This suggests that participants may attend to the primary dimension more than the secondary dimension, by which a change in the primary dimension becomes more noticeable than a change of equal size in the secondary dimension (see also Aha & Goldstone, 1992; Goldstone, 1994; Nosofsky, 1984).

The change of average proportion of correct rule and exception responses during training are shown in Fig 3-4 top left and top right panels (see also Erickson and Kruschke, 1998 Figure 3 and Figure 10). As can be seen from the figure, the training phase concluded with over 80% correct rule responses to rule-based exemplars and over 85% correct exception responses to exception exemplars. The results indicate that the participants had learned the experimenter-defined exemplar membership.

In addition, Fig 3-4, bottom right panel shows the proportion of rule responses to each exemplar. (Note that data from the bottom half in the original psychological space were rotated and combined with those in the top half of the category structure to generate this diagram, see also Erickson and Kruschke, 1998 Figure 4). As can be seen in Fig 3-4 bottom right panel, participants tended to given rule responses to both rule and exception critical transfer items.

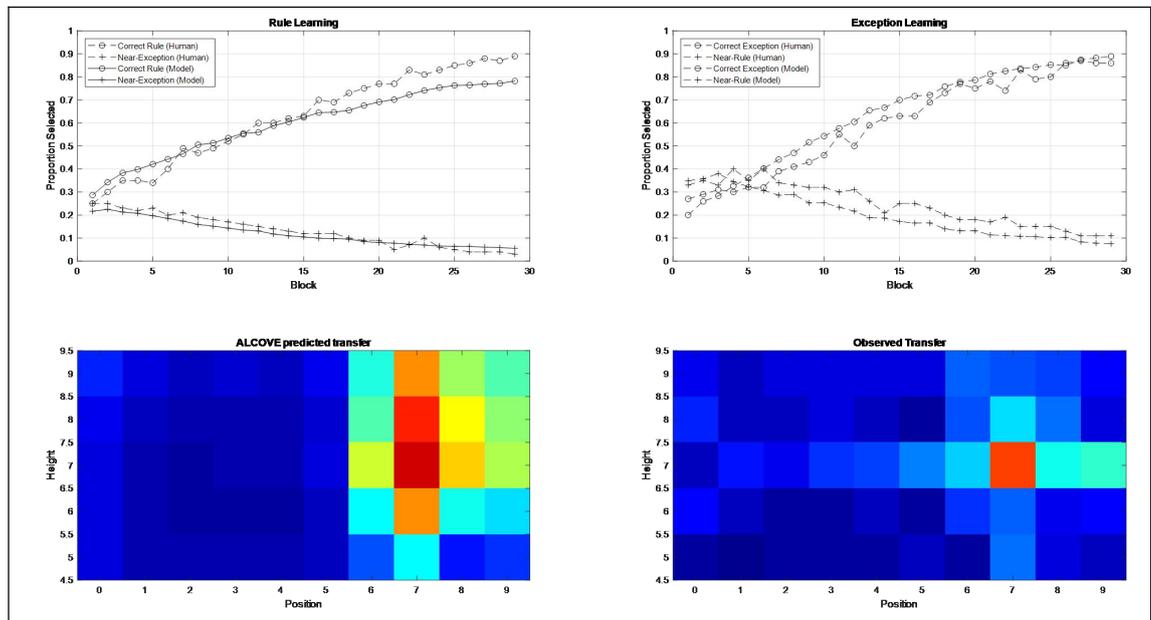


Fig 3-4. Top: Human data from Erickson and Kruschke (1998), experiment 1, and model data from the reimplementation of ALCOVE, showing the proportion of rule and exception responses as a function of learning blocks. Bottom: The ALCOVE reimplementation’s prediction of proportion of exception responses in the transfer phase (left) and observed human participants’ data in Erickson and Kruschke (1998) Experiment 1 (right). Cool colours represent low choice probability whereas warm colours represent high choice probability.

3.2.3.1 Method

For fitting the human data of the rule-plus-exception task, the implementation of ALCOVE follows the model assumptions described earlier in this chapter (see also Erickson & Kruschke, 1998; Kruschke, 1992). The same parameter settings as estimated by Erickson and Kruschke (1998) were used and the model was first trained over the course of 29 blocks of 14 trials, mimicking the experience of each human participant. In each block, stimuli were presented in a random ordered sequence. After training, the learned attention strengths and association weights were then applied to determine the response to the transfer stimuli in a random order. This replication included 100 simulated participants.

3.2.3.2 Results and Discussion

Fig 3-4 top panels show the model fitting to the average proportion of different responses to different stimuli across the 29 blocks observed in Erickson and Kruschke (1998). The replication here, as was done by Erickson and Kruschke (1998), has shown the power of the model in the prediction of human learning and generalisation. As can be seen in the figure, ALCOVE did a

good job in fitting human training data. But as a model based on homogeneous representation perspective, it fails to predict the extrapolation effects. The replicated model predicted over 43% exception response to the critical exception test items, whereas Erickson and Kruschke (1998) only reported 11%.

How does one explain, in a computational model, the observed training and transfer data? ALCOVE can account for rule-based behaviour when categories are perfectly separable along a single component, and stimuli in each category set are concentrated in the psychological space. But it does not fare well with the rule-plus-exception representation. However, the failure of ALCOVE was expected as this has been reported by Erickson and Kruschke (1998). According to Erickson and Kruschke (1998), ALCOVE does not include any mechanism that allows shifting attention between dimensions. Therefore, they introduced a gating mechanism in a hybrid architecture (which will be introduced in Chapter 6).

Although ALCOVE failed to predict rule-based extrapolation, it remains a very important theory in category learning. In the next section, its theoretical implications are discussed in depth.

3.3 Theoretical Implications

3.3.1 Maintenance and Perseveration of Learned Attention

As noted above, according to ALCOVE, category membership of an input stimulus is determined by its relative similarity to learned exemplars, and the relative similarity is determined by the distribution of dimensional attention strengths. Kruschke (1992) applied back-propagation to instantiate learning of exemplars and attention. In the previous sections the reimplementation has verified that ALCOVE is able to account for the relative difficulty and learning speed across different task types (see also Kruschke 1992; Love et al. 2004; Nosofsky et al., 1994a).

In addition to attention learning, the feed-forward network is able to reflect the maintenance and perseveration of dimensional attention strengths in categorisation. Because the back-propagation algorithm adjusts weights by reducing the difference between the actual output value and the expected output value, once the output matches the expected value, adjustment of association weights and attention strengths is not required. This mechanism therefore predicts that the learned attentional focus should perseverate into subsequent training, even if the dimension values or the category memberships change.

Kruschke (1996a) tested human learners' performance on four types of shift learning. Different from traditional designs of category learning, instead of changing the stimuli or outcomes when the relevance changes, all the stimuli and outcomes stay the same; only the mapping between them changes. Learners were initially trained on task Type II structure. After 22 8-trial blocks of training (i.e., each exemplar is randomly presented once in each block), trials were seamlessly shifted to one of the structures shown in Fig 3-5A. Results from human learners showed that the 'Reversal' and 'Relevant' shifts, which require one to fully or partially stay on original relevant dimensions, were shifted more quickly than the 'Irrelevant' and 'Compound' shifts, which require involving the originally irrelevant dimension.

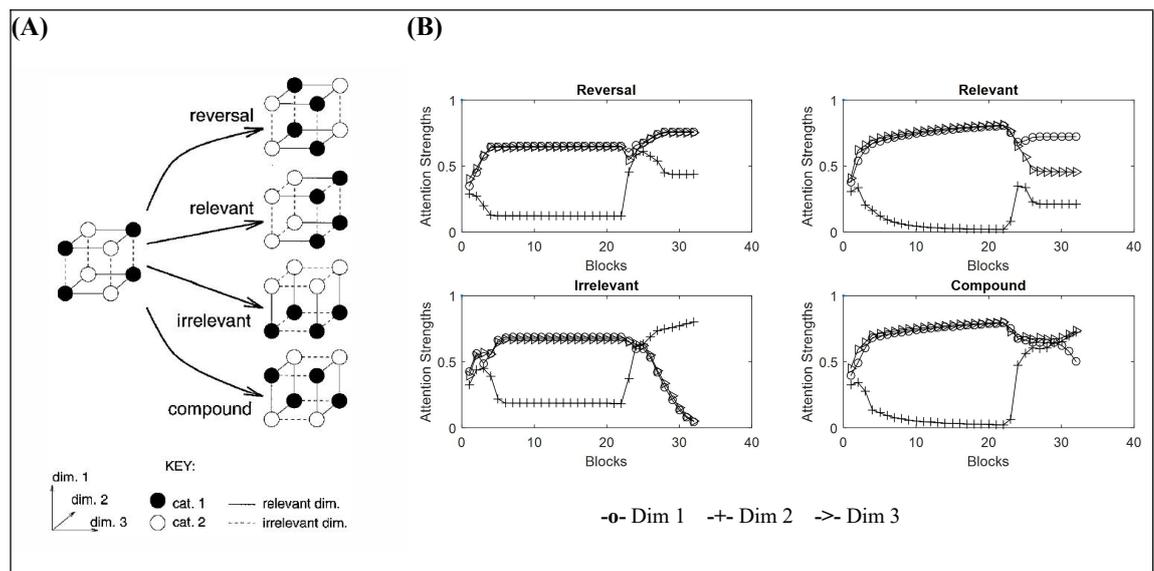


Fig 3-5. (A) Abstract structure of four types of shifts used in Kruschke (1996a) and (B) AMBRY predicted learned dimensional attention strengths as a function of training blocks.

An extended version of the ALCOVE (i.e., adding a category-to-response layer), called AMBRY (and introduced above), qualitatively fits the data. Fig 3-5B shows variation of dimensional attention strengths across training blocks in response to each shift learning condition predicted by AMBRY. Further studies of highlighting and blocking in category learning have found similar effects of attention perseveration – when trained on new associations involving previously highlighted cues, learning is faster; but, when trained on new associations involving previously blocked cues, learning is slower (e.g., Kruschke & Blair, 2000; Kruschke, 2005; see Kruschke, 2011, for a review). The feed-forward attention learning network based on the ALCOVE framework has been verified to be able to qualitatively account for these cases.

3.3.2 Relationship to Other Models

It is noteworthy that the power of the exemplar theory and attention learning framework is not only reflected in the performance of ALCOVE model and its extensions, but also reflected in its influence on other models of category learning. As mentioned in Chapter 2, most of the models based on the multiple representations assumption are more or less derived from the exemplar-based representation framework. SUSTAIN and ATRIUM both retain the attention learning principles. In SUSTAIN, the model recruits substructures to simplify a complex categorisation task (see Chapter 5 for details). Both solutions borrow from exemplar theory. ATRIUM is instantiated as a mixture-of-experts network. The sub-representation that requires attention to a single dimension is represented as a decision bound expert network, while the exemplar network is preserved as the expert to account for the cases which require attention to multiple dimensions. Selection of the appropriate sub-representation is mediated by the gating network (see Chapter 6 for details). In addition, the exemplar-based representation has been also considered by the neurobiological theory (Ashby & Rosedahl, 2017). Ashby and Rosedahl (2017) argued that the procedural learning system of their COVIS model can be interpreted as a neural version of the exemplar theory. This is because while COVIS does not suggest attentional processes in the procedural learning system, it does argue that the procedural learning system is designed to integrate perceptual information of input stimulus.

3.3.3 Attention and Multiple Representations

As can be seen in Section 3.2.2 ALCOVE can explain the rule-based category structures (task Types I and II) and the information-integration category structures (task Types III-VI) in different ways by assigning different attention strengths on dimensions. This seems not to conflict with some multiple representation accounts, such as COVIS. But, unlike COVIS, that argues that information-integration category learning cannot require attentional processes, ALCOVE suggests that representation of information-integration categories is formed on the basis of attention learning. However, both mechanisms implement the perceptual integration of stimulus input information. Therefore, in cases with homogeneous representations it seems not necessary to contrast these accounts.

ALCOVE's feed-forward attention learning framework is designed on the basis of the traditional assumption that the representation involved in a categorisation task is homogeneous. However, sometimes, representation of a categorisation task can be heterogeneous. Section 3.2.3,

for example, has shown that the ALCOVE is unable to predict rule-based extrapolation after learning rule-plus-exception categories (see also Denton et al., 2008).

As mentioned in Chapter 2, Aha and Goldstone (1992) provided a set of highly flexible category structures (see Fig 2-7A). They sampled training stimuli from two distinct category structures in a two-dimensional stimulus space, each of which was bisected by its own uniquely oriented boundary. When extended further from their cluster, the boundaries dictated opposite classifications for the same test items (see Fig 2-7B). Aha and Goldstone found that participants classified transfer stimuli using the closest partial boundary. This result reveals that selection of categorisation behaviour is dependent on differences between subsets of stimuli.

Another paradigm which has shown evidence of heterogeneity of category representation is the task partitioning paradigm (e.g., Lewandowsky et al., 2006; Little & Lewandowsky, 2009; Yang & Lewandowsky, 2003; 2004). In task partitioning, a normatively irrelevant context cue (e.g., the background colour in which a stimulus is presented with) reliably signals the presence of a local regularity in the task (e.g., the presence of a partial category boundary). When task partitioning occurs, people simplify a complex task by decomposing it into separate local solutions.

Unlike the traditional interpretation of category representation, task partitioning requires individuals to use multiple sub-representations strategically to decompose a complex, heterogeneous categorisation tasks into simpler ones that place reduced demands on the limited capacity of attention. But, obviously, the dimensional attention learning mechanism is unable to meet this requirement.

Erickson, (2008) argued that the sub-representations in the heterogeneous categorisation tasks are determined by the distribution of dimensional attention strengths. Success in learning of the categorisation tasks is based upon the allocation of learned dimensional attention strengths in response to task demands. This mechanism is analogous to the conception of task switching. Task switching allows one to efficiently adapt to different situations. In the structure provided by Erickson (2008), for some cases, attention must be assigned to a set of dimensions, but not a single dimension. In other words, it requires a mechanism to mediate shifts between representations that are determined by various distributions of dimensional attention strengths. In addition, in Erickson's task, contextual cues are also given to indicate which representation to use. In task switching, it is assumed that an executive control mechanism is responsible for mediating task switching. If the contextual cue repeats, executive control does nothing, and the stimulus is processed in line with the task set from previous trial. However, if the cue alternates, executive control inhibits the previous task set and excites the indicated one (Shallice et al., 2008, see also Kiesel for a review).

In sum, although, ALCOVE does not suit multiple representations, especially, the heterogeneous category representations, it remains necessary to give insight into the exemplar theory. This is because, on one hand, most of the multiple representation models are more or less derived from it, and, on the other hand, attention should play an important role in category learning.

Chapter 4.

Case Study I: SUSTAIN and Multiple Representations

4.1 Introduction

Categories with family-resemblance structure are learned well by models that represent prototypes. A prototype can be thought of as the central tendency of a category or as the best representative of a family, in the sense of having the greatest number of attributes in common with other exemplars of the category and the fewest number of attributes in common with exemplars of contrasting categories. Unlike rule-based theory and exemplar theory, prototype theory suggests that category membership of an exemplar is determined by its similarity to a few central exemplars of that category. For example, when asked to give an example of the category of ‘mammal’, chimpanzee should be cited more frequently than bat. Traditional prototype models are inflexible in that they treat the structure of each category as predetermined. However, in the real world, some categories are comprised of highly nonlinear substructures. For example, spoons could be small and made of porcelain or large and made of steel. For the category *spoon*, there is not a characteristic weighting of the dimensions of material and size; rather, there are two distinct substructures that contain opposite values on these two dimensions. But, meanwhile, the structure of some categories can be simpler. However, prototype models that assume that categories are always represented by one node (e.g., Minsky & Papert, 1969; Posner & Keele, 1968; Rosenblatt, 1958) cannot learn categories rich substructures, while others complex models with more complex representations (e.g., Rumelhart et al., 1986) may perform poorly with

simpler categories. For learning simple categories, overly complex models will tend to generalise poorly by overfitting the training data. This tradeoff between data fitting and validation is termed the bias-variance dilemma (Geman et al., 1992). SUSTAIN (Supervised and Unsupervised STRatified Adaptive Incremental Network) is a kind of prototype-based hybrid network model developed to provide a solution to this dilemma by achieving a balance between the two extremes (Love et al., 2004).

Category learning in humans begins with simple solutions, and changes in response to ‘surprising events’ (Ashby et al., 1998; Nosofsky et al., 1994a; 1998). At the same time, one reasonable intuition is that similar items should be clustered together in memory (Anderson, 1991; Love et al., 2004). SUSTAIN embodies these ideas. SUSTAIN recruits new clusters in response to surprising events. The cluster recruitment processes are driven by situational goals. The original version of SUSTAIN is comprised of both supervised and unsupervised learning mechanisms. In unsupervised learning, items that are dissimilar from existing prototypes result in a new cluster being recruited to encode the item. However, in the supervised category learning situation, as the goal is to adaptively learn each stimulus’s category membership, items are recruited as new clusters when a surprising error results. Therefore, the more complex the task, the more surprising errors occur during learning, and, thus, the more clusters will be recruited.

In addition, SUSTAIN also embodies a selective attention learning mechanism (e.g., Erickson & Kruschke, 1998; Kruschke, 1992; 1996a; Nosofsky, 1986) that can learn to emphasise dimensional relevance. The selective attention learning mechanism allows SUSTAIN to display rule-like behaviour when learning a simple rule. This issue will be discussed later in more detail.

In this Chapter, the SUSTAIN model will be formally introduced. The Chapter presents a reimplemention Love et al.’s (2004) work fitting Shepard et al.’s (1961) six problems, together with further testing of the SUSTAIN model’s behaviour in the information-integration (II), hybrid and rule-plus-exception category learning tasks. The focus in this Chapter is only on the supervised classification learning functions of SUSTAIN. The inference learning and unsupervised learning functions will not be discussed.

4.2 Reimplementing SUSTAIN

4.2.1 Model Description

The basic structure of the supervised learning network of SUSTAIN is illustrated in Fig 4-1. Starting at the input layer of the network, perceptual information is transformed into a set of row vectors that is organised along a set of dimensions. The illustrative example shown in the figure consists of two dimensions. Attention weights are learned for each dimension. The weights determine dimensional relevance. The hidden units are a set of substructures that are each associated with a category. Each input stimulus is assigned to a category according to the most activated hidden unit. When the assignment is incorrect, a new cluster is recruited to represent the current instance.

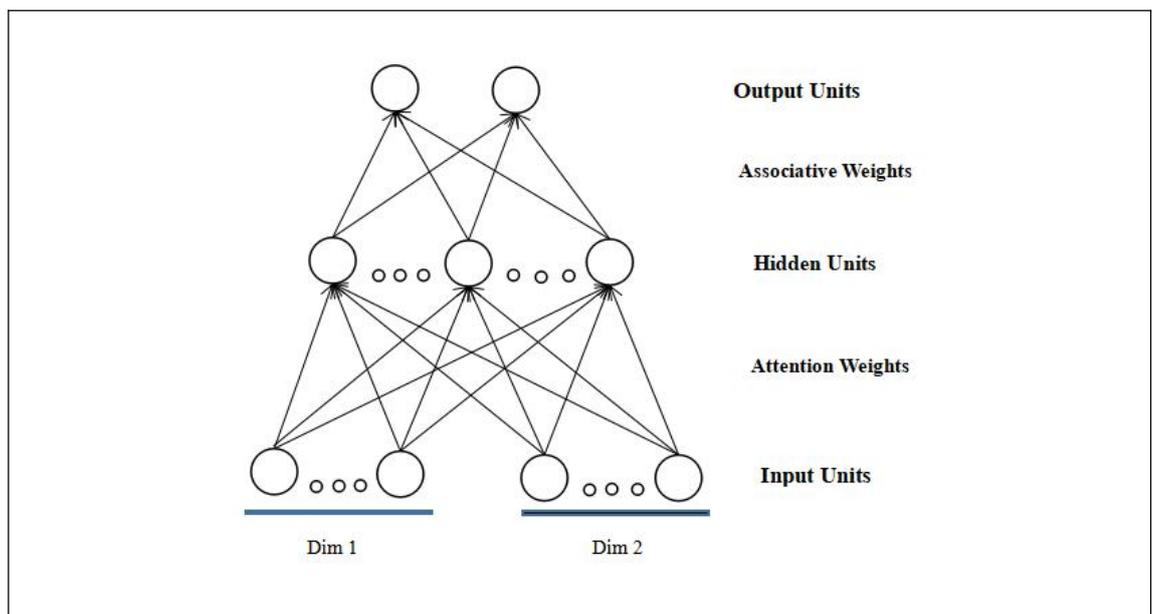


Fig 4-1. An illustrative representation of architecture of supervised classification learning network of SUSTAIN model. (Dim = dimension)

Three Principles Supervised learning in SUSTAIN follows three simple principles. First, category learning begins with simple solutions. This principle is in line with the multiple systems theory of category learning (e.g., Ashby et al., 1998; Nosofsky et al., 1994a; Sloutsky, 2010). Second, similar items cluster together in memory (i.e., hidden units) throughout learning. In learning to classify stimuli as members of the category birds or mammals, for example,

SUSTAIN clusters similar items together. Instances similar to a bird substructure *sparrow* could cluster together and form a *sparrow* cluster instead of leaving separate traces in memory. Although similarity drives clustering, clustering also drives similarity as attention shifts to stimulus dimensions that yield consistent matches across clusters. Third, selection of hidden substructures is competitive. Given the hidden units are competing explanations that attempt to explain the input information, the strength of the response from the most similar cluster is attenuated in the presence of other clusters that are, to some extent, similar to the current instance.

Selective Attention Mechanism Each cluster consists of a receptive field for each given stimulus dimension, which is centered at the recruited item's position along that dimension. The influence of the position of a receptive field is determined by how much attention is assigned to the stimulus dimension. All the receptive fields for one stimulus dimension share a common attentional weight. The attentional weights change as a result of learning. The attentional leaning mechanism is similar to that of exemplar-based network models (Kruschke, 1992; 1996a; 1996b). Dimensions that provide consistent information at the cluster level receive greater attention.

Based on the selective attention mechanism, the activation of a hidden unit is mathematically given by

$$a_n^{hidden} = \frac{\sum_i (\lambda_i)^r \exp(-.5\lambda_i \sum_k |I^{i,k} - H_n^{i,k}|)}{\sum_i (\lambda_i)^r} \quad (4.1)$$

where λ_i is the attentional weight on dimension i , r is a nonnegative constant called the *attentional focus* parameter, I is the input representation and H_n is cluster n 's position on the receptive fields of dimension i .

Cluster Recruitment and Competition The hidden units compete to respond to input information and in turn inhibit one another. For a winning unit a_{winner}^{hidden} ,

$$a_{winner}^{hidden} = \frac{(a_{winner}^{hidden})^\beta}{\sum_n (a_n^{hidden})^\beta} a_{winner}^{hidden}, \text{ then, for all other hidden units } a_n^{hidden} = 0. \quad (4.2)$$

where β is another nonnegative constant called the *competition* (or lateral inhibition) parameter. The competition parameter is designed to regulate competition among hidden units. When β is large, the winner is only weakly inhibited. Clusters other than the winner are not selected and have their activation value set to zero. Note that only the winning cluster determines the activation of the output units.

After responding, feedback is provided to SUSTAIN. Like attentional learning models, supervised learning in SUSTAIN is determined by a gradient error reduction mechanism. The feedback is coded as humble teacher values, t_r , given to each output unit. The teaching values are defined as $t_r = +1$, if the stimulus is a member of category r , and $t_r = 0$, otherwise. Basically, the model is not penalised for predicting the correct response more strongly than is necessary.

A new cluster is recruited if the winning cluster predicts an incorrect response. The recruitment of a cluster follows the rule that if t_r does not equal 1 for the most activated output unit, then, recruit a new cluster. In other words, the output unit representing the correct nominal value must be the most activated. When a new cluster is recruited, the hidden units' activations and outputs are recalculated. The new cluster becomes the winner because it is the most activated cluster. Only the winning cluster updates the attention weights and associative weights. (The learning rules will not be discussed in detail, see Love et al., 2004, for more detail.)

4.2.2 Modelling Shepard et al.'s Six Problems

4.2.2.1 Rationale

As mentioned in Chapter 2, as one of the most classic studies in human category learning, Shepard et al.'s (1961) six classification problems have provided a very challenging data set for category learning models. Nosofsky et al. (1994a) replicated the Shepard et al. (1961) experiment, and exemplar-based network models ALCOVE and the RULEX model have successfully fit the data. ALCOVE solved the six problems via the adjustment of attentional weights on each dimension. For example, for the simple rule, attention to relevant dimension would be much greater than others, while, for Type VI problem, attention weights on three dimensions would show little differences. In contrast, the logic of RULEX is that the learner begins with simple rules, and gradually moves on to imperfect rules with exceptions. This logic

is in line with the differences in learning speed in various classification problems. However, SUSTAIN is a hybrid network that incorporates a selective attention mechanism and the logic of development from simple solution to complex solution (via cluster recruitment). Therefore, not surprisingly, Love et al. (2004) showed that SUSTAIN was also able to fit this data set.

4.2.2.2 Method and Results

The reimplementing of SUSTAIN follows the model description and parameter settings (see Table 4-1) provided by Love et al. (2004) and summarised above. The results are shown in Fig 4-2. The procedure used to simulate SUSTAIN mimicked the procedure used to collect data from the human participants. As can be seen in Fig 4-2A, SUSTAIN's fit of Nosofsky et al. (1994) is very good. In particular, SUSTAIN correctly predicts that Type I is learned faster than Type II, which is learned faster than Types III–V, which are learned faster than Type VI.

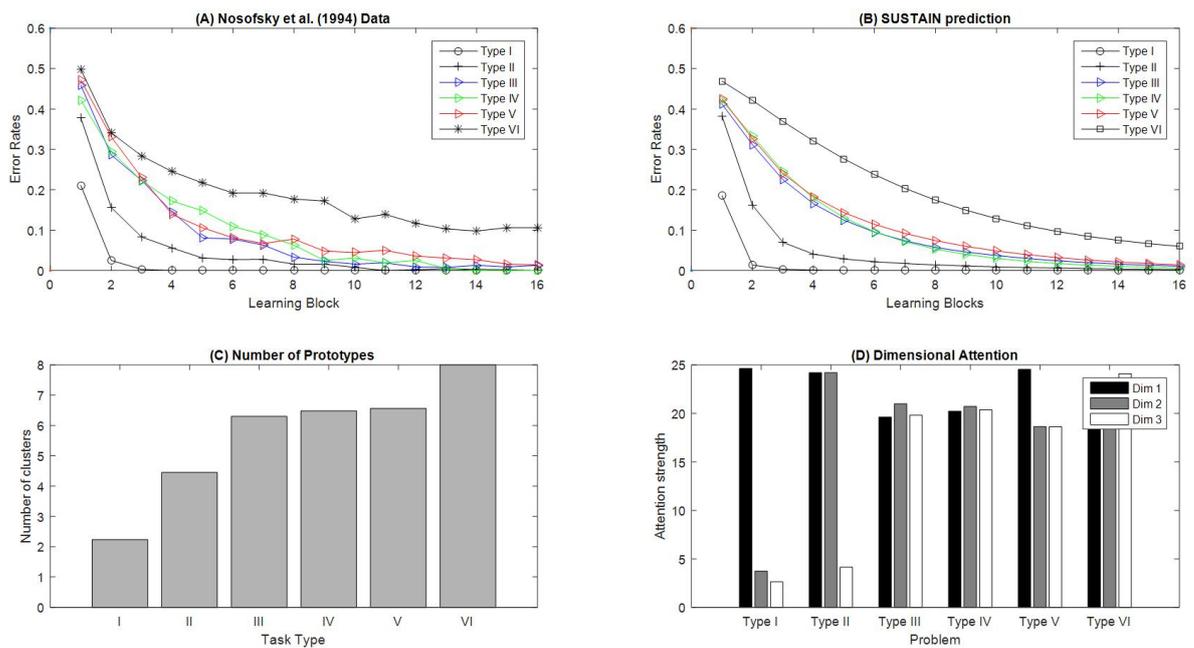


Fig 4-2. SUSTAIN's fit of Nosofsky et al. (1994a) data (A and B), number of hidden prototypes recruited throughout training (C), and assignment of attentional strength to each dimension (D).

As the selective attention mechanism and cluster recruitment mechanism are the core components of SUSTAIN, Fig 4-2B and Fig 4-2C, respectively, show how complex substructures and dimensional attention strength are learned through training. Basically, the learned dimensional attention strengths reflect the regularities of the different problems. For

instance, Type I problems require attention almost exclusively to the first dimension, while attention to two dimensions are required for solving Type II problem. Type VI has no regularities that can be exploited, and thus it requires attention to all dimensions (i.e., to memorise each item). At the same time, the number of clusters SUSTAIN recruits varies with problem difficulty. For instance, the solution commonly adopted for the simple rule (Type I) problem involves recruiting one cluster for each category. In contrast, the Type VI problem has no regularities, forcing SUSTAIN to memorise each stimulus (i.e., the model devotes a cluster to each item).

Table 4-1.
Function and best fitting values of parameters for reimplementing SUSTAIN

| Parameter | Function | Shepard et al. (1961) |
|------------------|----------------------|------------------------------|
| ϕ | Decision consistency | 16.92 |
| r | Attentional focus | 9.01 |
| β | Lateral inhibition | 1.25 |
| λ | Learning rate | 0.09 |

4.2.2.3 Discussion

The reimplementation of SUSTAIN following Love et al. (2004) basically shows the power of this hybrid network model at fitting the classic category learning data set. Although the Shepard et al. (1961) paradigm includes a variety of category structures, it has an obvious limitation that each stimulus in the study only involves binary-valued, but not continuous, dimensional input. In addition, the Shepard paradigm, as mentioned in Chapter 2, does not involve a heterogeneous category structure which needs selection between multiple representations. As shown in Fig 4-2, during learning the average numbers of substructures recruited for task Types III-V (6.11, 6.18, 6.25) are significantly fewer than the average number of substructures recruited for task Type VI (mean = 8). This is because SUSTAIN, sometimes, applies rule-plus-exception like strategies (i.e., focus on one of three dimensions more than the others). In the sense, it is possible that SUSTAIN can show better performance on rule-plus-exception category learning than ALCOVE. But, as SUSTAIN, sometimes, also applies the II strategy, there is another possibility – that SUSTAIN finally develops an

exemplar-based representation for the rule-plus-exception category structure. The next section considers how SUSTAIN may account for the formation of multiple representations in category learning.

4.3 Simulation Studies

4.3.1 Simulation I: II Category Learning

4.3.1.1 Rationale

As mentioned earlier, one advantage of SUSTAIN is its flexibility in learning complex category structures which may be comprised of multiple substructures. The reimplementations of SUSTAIN in the modelling of Shepard's six problems has shown the flexible power of the model to account for various deterministic categories. As an incremental adaption architecture, SUSTAIN can be very effective in mastering a wide range of learning problems because it can adapt its complexity in response to the current problem.

However, although Love et al. (2004) argued that SUSTAIN can be used to simulate category learning with continuous stimulus dimensions, to my knowledge, no attempt has been made to simulate learning II categories. Unlike other models of multiple representations, SUSTAIN does not follow a modular architecture. Instead, SUSTAIN suggests that category learning is determined by the processes of generating and selection between multiple substructures. These processes depend heavily on the dimensional attention allocation mechanism. As the variance between attention strengths decreases, more and more substructures will be generated. As can be seen in the simulation of Shepard et al.'s (1961) deterministic task, for a category structure with no regularity, the model comes to learn to use an exemplar-like strategy. The model learns to integrate dimensional information by assigning a similar level of attention strength to each dimension. Thus, in principle, the model can learn II categories by recruiting more prototypes than for RB categories. In this section, and for illustrative purposes, the reimplementations of SUSTAIN is therefore applied to simulate learning bivariate-normally distributed RB and II categories.

4.3.1.2 Category Structures

The category structures of the RB and II categories used in this simulation are shown in Fig 4-3. The attributes of simulated category stimuli mimic the features of Gabor patches introduced by Ashby and Gott (1988). Thus, the two dimensions are bar frequency and orientation. For training the network, a total of 300 stimuli were drawn from each of two bivariate normal distributions in each task. The category distribution parameters (i.e., mean and variance on each dimension, and covariances between dimensions) were as follows. For the 1D rule-based categories, category A had a bar frequency mean of 40 and an orientation mean of 50, whereas category B had means on these two dimensions of 60 and 50, respectively. In both categories, the variance on dimension bar frequency was 10 and orientation was 200, and the covariance between dimensions was 0. For II categories, category A had means on the two dimensions of 60 and 40, whereas category B had means of 40 and 60, respectively. The variance on each dimension was 167.59, and the covariance between dimensions was 151.24.

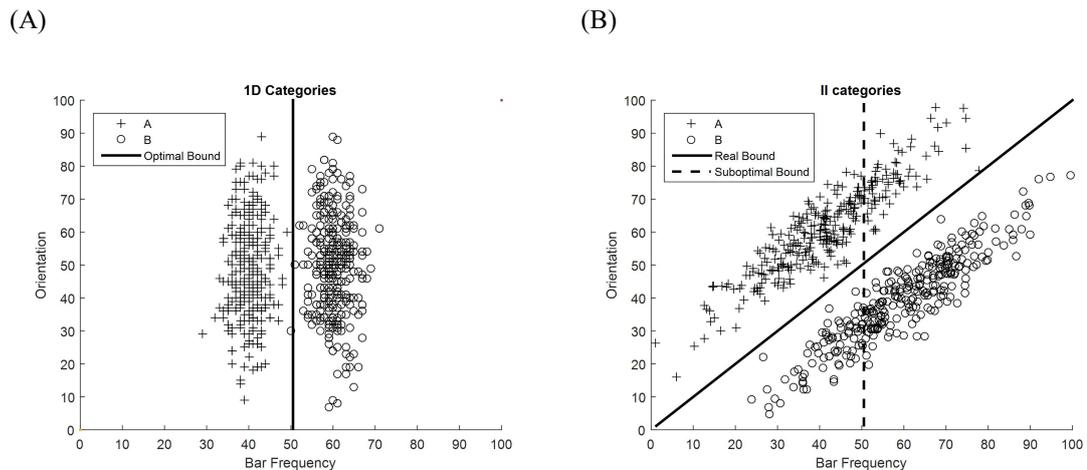


Fig 4-3. Schematic representation of one dimensional rule-based (1D) and information integration (II) category structures used in the simulation. Dot marks (plus and circles) represent the 2D coordinates of stimuli (e.g., disks varying in spatial frequency and orientation). The solid black line represents the optimal boundary, the dashed block line represents a suboptimal rule-based boundary.

4.3.1.3 Implementation Issues

The model implementation followed the original, supervised learning version of SUSTAIN described above with some modifications. These modifications were necessary because, as mentioned earlier, the model was originally designed to account for deterministic, nominal categorisation stimuli (Love et al., 2004). To account for categories with bivariate normal

distributions, it was necessary to redefine the input representation and hypothetical maximal distance between input stimulus and hidden units in the psychological space.

In the original version of SUSTAIN, stimuli were represented as vector frames where the dimensionality of the vector was identical to the dimensionality of the stimuli. For example, a four-dimensional, binary-valued stimulus can be thought of as a four-character string in which each character represents a stimulus dimension. The first character could denote the size dimension with 1 indicating small and 2 indicating large. However, in a continuous-valued category structure, each stimulus can be thought of as a point in psychological space that activates nearby exemplars strongly and distant exemplars more weakly. With nominal stimulus dimensions, the distance between the stimulus and hidden units' position on each dimension ranges between 0 and 1. However, in this simulation, as the psychological space is represented as a 50 x 50 grid, the maximal distance in this space would be 49, which would lead to very rapid changes in the strength of dimensional attention. To deal with this, the values of distances were renormalised to a [0, 1] scale. This was done by dividing each value by the possible maximal distance.

4.3.1.4 Method

As SUSTAIN has only four important parameters, all these parameters were considered. In a preliminary parameter search phase, it was found that the learning rate parameter (λ) does not influence the model performance. In addition, the decision consistency parameter (ϕ) is just a scaling constant representing the level of decisiveness for the final output of the system. As mentioned earlier, the extra substructure recruitment of SUSTAIN is highly influenced by the selective attention mechanism, whereas the final output of the model results from competition between substructures. Thus, for carrying out the evaluation, what is needed is to search in a two-dimensional parameter space formed by the attentional focus parameter r and the competition parameter β , because these two parameters are critical to substructure recruitment and competition. In doing so, it was chosen to fix the value of ϕ at 17.00 (the same arbitrarily value chosen in the original Love et al. (2004) paper) and λ at 0.03 (less than the value of 0.09 used for binary-valued feature-based classification by Love et al. (2004), but sufficient for probabilistic feature-based categories). As the purpose of this simulation was to illustrate how SUSTAIN can learn different bivariate-normally distributed categories, the focus was on model

performance based on some specific regions in the parameter space rather than across the full space (see Table 4-2). This simulation consisted of 100 Monte Carlo simulation steps (random sampling in the specified parameter space). Each simulation step included 15 blocks of 600 trials training on RB categories and II categories. The stimulus ordering and parameter values across tasks were reinitialised at the beginning of each block.

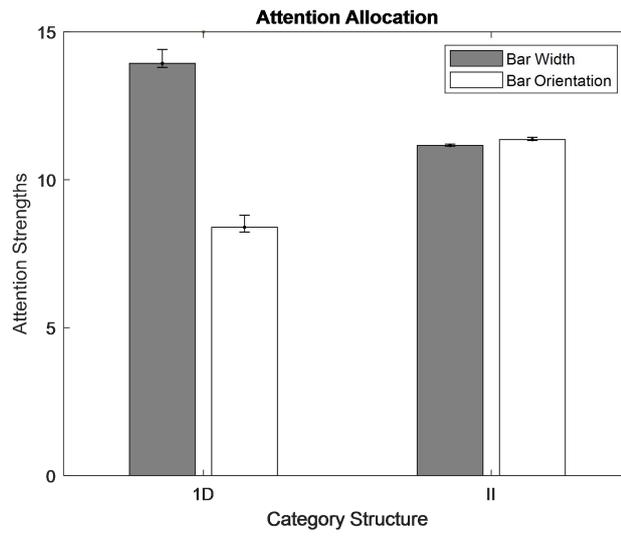
Table 4-2.
The regions in the parameter space specified for simulations

| Parameter | Simulation | |
|-----------|------------------------|----------------------|
| | Bivariate Distribution | Uniform Distribution |
| ϕ | 17.00 | 19.00 |
| r | [2.00, 4.00] | [2.00, 4.00] |
| β | [8.00, 10.00] | [8.00, 10.00] |
| λ | 0.03 | 0.18 |

4.3.1.5 Results

This simulation study showed that SUSTAIN can form representations of bivariate-normally distributed RB and II category structures in different ways. In particular, for dimensional attention allocation, not surprisingly, learning II categories produces similar attention strengths on the two dimensions, whereas learning RB categories lead to greater strength on the bar frequency dimension (x-axis) than on the orientation dimension (y-axis) (see Fig 4-4A). In addition, as shown in Fig 4-4B (histogram), SUSTAIN needs to recruit more substructures in learning II categories (mean = 13.96) than RB categories (mean = 5.58). Fig 4-5 (upper panel) shows the result of searching in the two-dimensional parameter space. It is obvious that parameter values resulting in the best model performance for both tasks falls in the area where r is small and β is great. The results suggest that SUSTAIN can provide a good account of the dissociation between representations of RB and II category learning.

(A)



(B)

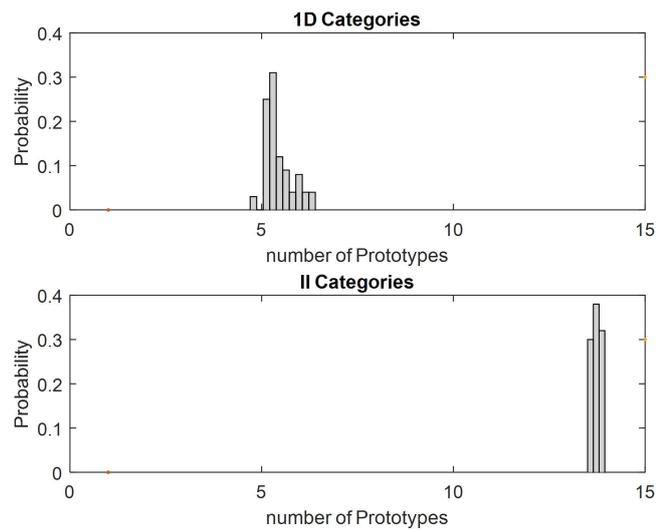
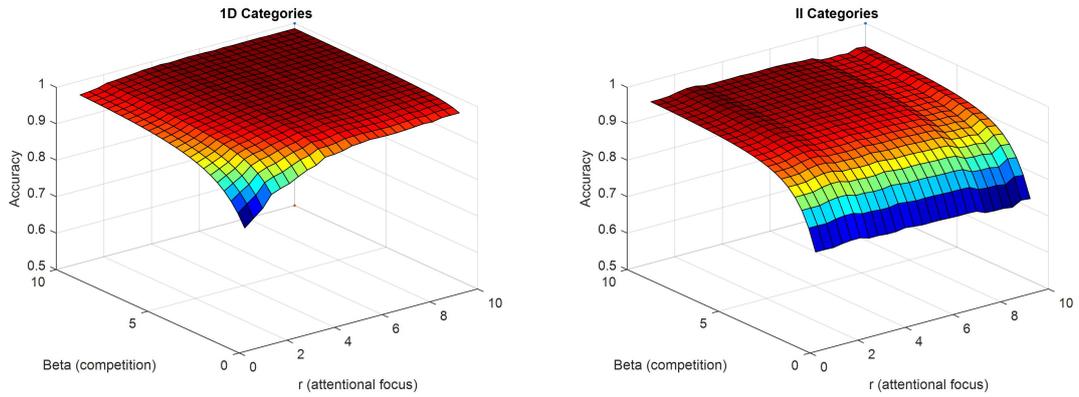


Fig 4-4. (A) Mean attention strength on each dimension for 1D and II category structures produced by SUSTAIN. Error bars represent maximum and minimum values produced by the model. (B) Histogram of distribution of probabilities the model recruiting certain number of prototypes as a result of learning.

Sampling From Bivariate Normal Distribution



Sampling From Uniform Distribution

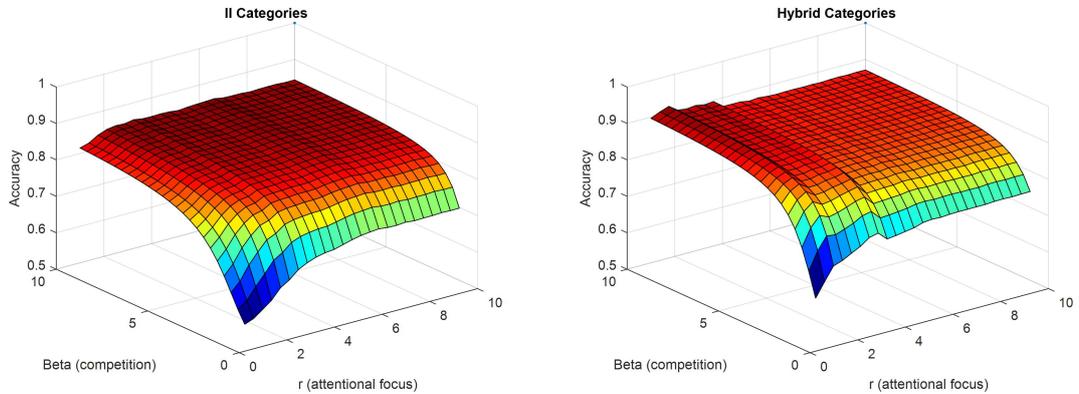


Fig 4-5. Effects on predicted accuracy of varying attentional focus parameter (r) and competition parameter (β) for (upper panel): learning bivariate normal 1D RB and II, and (lower panel): uniformly distributed II and hybrid category structures.

4.3.1.6 Discussion

While SUSTAIN can learn categories based on bivariate normal distribution, the representation it produces during learning seems different from those it produces when learning binary-valued 1D categories. For example, as can be seen in Fig 4-2, only two clusters (mean = 2.05) were recruited for 1D category representation, and the attention strength on the diagnostic dimension is much greater (mean = 25.07) than the non-diagnostic dimensions (both means are less than 5.00). However, in the bivariate-normally distributed categories, more clusters were recruited. This may be because the difference between the diagnostic and non-diagnostic dimensions is large (mean_{diagnostic} = 14.08 and mean_{non-diagnostic} = 8.43), but the strength on the non-diagnostic dimension remains too high to be ignored during learning. This prediction is inconsistent with some other machine learning-based attention models such as ALCOVE and

ATRIUM which suggest that learning should result in the attention on the non-diagnostic being decreased so that they are neglected after category learning.

4.3.2 Simulation II: Hybrid Category Learning

4.3.2.1 Rationale

Compared with exemplar-based network models, SUSTAIN is more sensitive to distributional properties of category. As a result, SUSTAIN appears to require different patterns of parameter settings for categories with bivariate normal distributions and categories with uniform distributions, but how does it fare with hybrid category structures? As can be seen from Fig 4-6, the distribution of members of uniformly distributed II and hybrid categories is very discrete. This type of category structure will encourage SUSTAIN to recruit more substructures than for categories of bivariate distributions. At the same time, Ashby and Crossley (2010) reported that human participants tended to use a rule-like strategy (i.e., greater attention strength on bar frequency) to solve the problem. It will not be a problem for SUSTAIN to assign greater strength to bar frequency, because, as mentioned in Chapter 2, all exemplars of category A in the hybrid structure can be perfectly categorised by the simple rule. The problem here, however, is how does SUSTAIN form the representation of category B? As Ashby and Crossley (2010) proposed, the use of simple rules may inhibit the II strategy, meaning that the exemplars of category B in the left corner of Fig 4-5B may receive more category A responses than the right side. This simulation is intended to test this hypothesis, and give further insight into the SUSTAIN model.

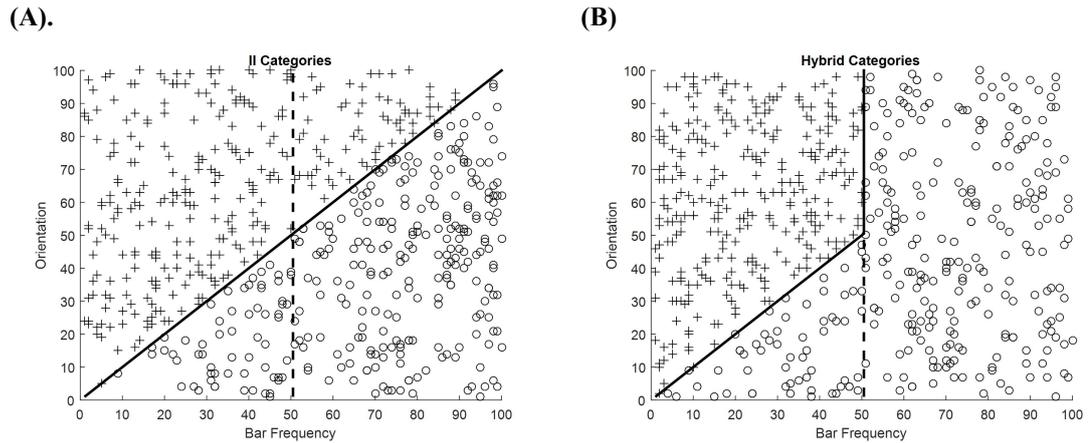


Fig 4-6. Schematic representation of uniformly distributed II (A) and hybrid (B) category structures used in the simulation. Dot marks (plus and circles) represent the 2D coordinates of stimuli (e.g., disks varying in sparial frequency and orientation). Tthe solid black line represents the optimal boundary, the dashed block line represents a suboptimal boundary.

4.3.2.2 Category Structure

The category structures used for this simulation were uniformly distributed II and hybrid categories as shown in Fig 4-4. To create exemplars of hybrid categories 300 points were randomly sampled from a uniform distribution defined over the A region illustrated Fig. 4-7B. Another 300 stimuli were randomly sampled from the uniform distribution defined over the B region. The result of this process was 600 ordered pairs of numbers (x, y), with each value ranging between 0 and 100. For the II categories, 600 random samples were drawn from the uniform distribution defined over the entire 0–100 space. Each (x, y) pair was given a category label according to the following rules: if $x < y$, then the stimulus was in category A, otherwise the stimulus was in category B.

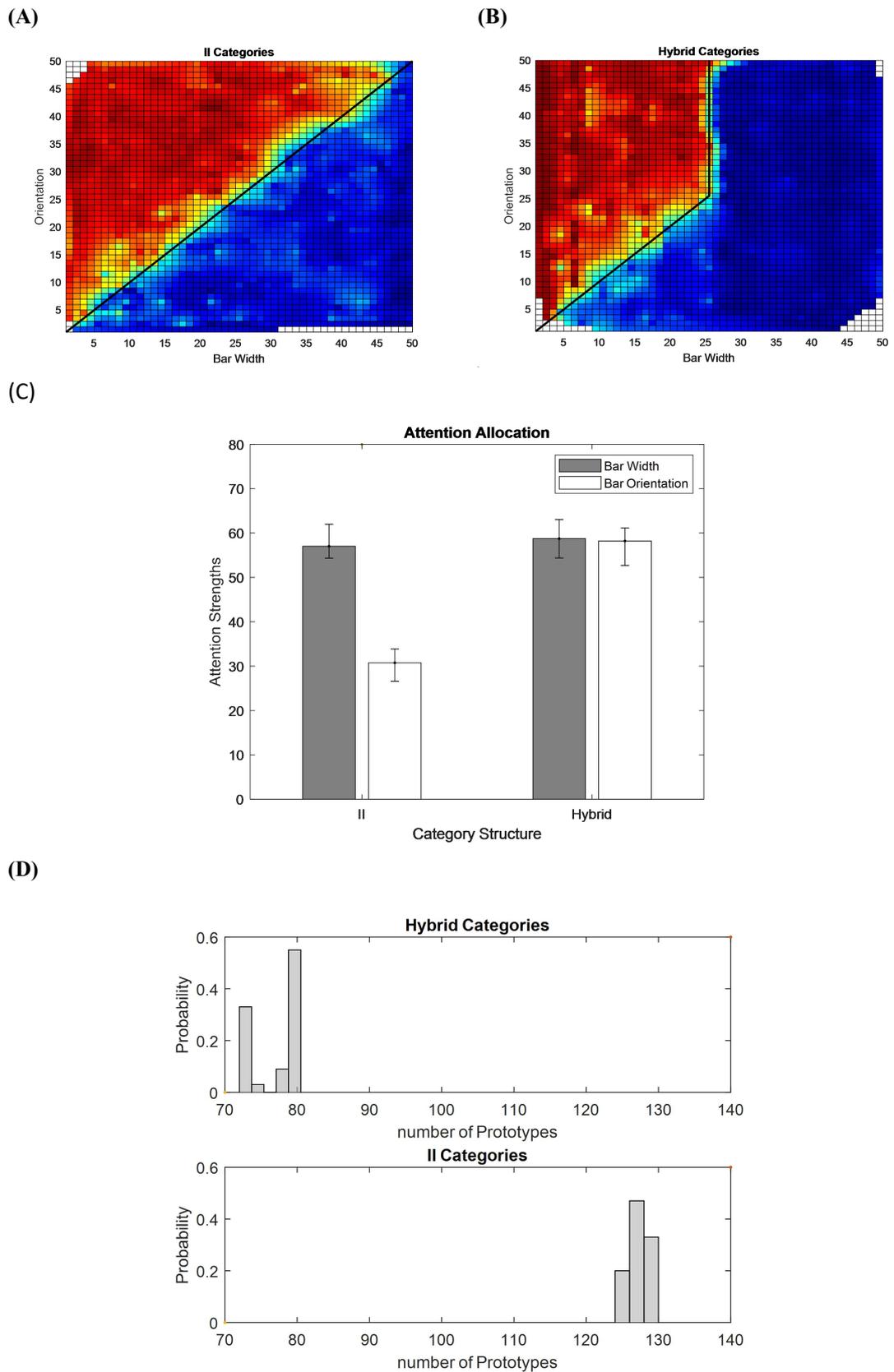


Fig 4-7. Top: Response A probabilities predicted by SUSTAIN for II (A) and hybrid (B) category learning (Blue (cool) colours represent low probabilities, whereas red (warm) colours represent high probabilities). (C): Mean attention strength on each dimension for uniformly distributed II and hybrid category structures produced by SUSTAIN. Error bars are maximum and minimum values. (D): Histogram of distribution of probabilities the model recruiting certain number of prototypes as a result of learning.

4.3.2.3 Method

Preliminary simulations found that SUSTAIN's learning on categories with a uniform distribution requires greater values on learning rate and decision consistency than learning on categories with bivariate normal distributions. Thus, for this simulation the value of the decision consistency parameter (ϕ) was fixed at 19.00, while the learning rate (λ) was fixed at 0.18. The simulation process was identical to that of section 4.3.1 (for specified parameter regions see Table 4-2), except that, in addition to the training session each simulation step, the learned model settings (e.g., dimensional attention, weights between hidden nodes and category response nodes, and learned prototypes) were used to run another 600 trial transfer phase with same set of stimuli and fixed random stimulus ordering.

4.3.2.4 Results

Fig 4-7A and Fig 4-7B show the choice probability on each stimulus averaged across the 100 simulation steps for II and hybrid category learning. As can be seen in the contour plot, SUSTAIN is robust in learning these category structures. The diagonal and hybrid boundaries are both captured by the model. These predictions are, to some extent, not in line with Ashby and Crossley's (2010) findings that human participants have difficulty in capturing the hybrid category structure (see also Paul and Ashby, 2013).

According to Fig 4-7C and Fig 4-7D, SUSTAIN learned to form representations of II and hybrid categories in different ways. The learned attention strength on each dimension across tasks indicate that the learning strategy of hybrid categories is similar to that of RB categories in that attention strength on bar frequency is much greater than on orientation (see Fig 4-7C). This is different from the II category learning. In addition, as expected, more substructures are recruited at the end of training on II categories than on hybrid categories.

4.3.2.5 Discussion

Although, according to Figs 4-7C and 4-7D, SUSTAIN has formed different representations for II and hybrid categories. It is still problematic. As mentioned above, in learning large-scale probabilistic datasets, SUSTAIN tends to be unable to neglect the non-diagnostic feature in an RB task. Ashby and Crossley (2010) also found that their participants tended to neglect the non-diagnostic dimension in learning hybrid categories. However, in SUSTAIN's representation of a hybrid category, though, attention strength is lower on non-diagnostic dimension, the

influence of the non-diagnostic dimension cannot be neglected after training. This property of SUSTAIN led to a large number of clusters being recruited during category learning. As a result, SUSTAIN did not show any difference for the left corner exemplars and right side exemplars in category B. Instead of inhibiting the alternative representation, SUSTAIN formed a hybrid category representation in a stimulus-dependent way (Erickson, 2008). This may be because in SUSTAIN learning is interpreted as the process of deconstruction based on the similarity between an input stimulus and stored substructures. This deconstruction process is demanding. Once the stored prototypes cannot meet the task requirements, the model continues to deconstruct the categories. This mechanism is computationally reasonable, but it is cognitively expensive, and may not be the way humans learn continuous-valued categories.

As described above, the attentional focus parameter addresses the importance of lateral inhibition between dimensional attention strengths, while competition parameter addresses the effect of lateral inhibition between internally recruited clusters. Fig 4-5 (lower panel) shows the effects of varying the competition parameter (β) and attentional focus parameter (r) on model performance. The figure shows that better performance of the model appears to require greater values on the competition parameter, but this effect may disappear if it exceeds a certain range. This is intriguing, because it appears to suggest that categorical decisions may depend heavily on lateral inhibitory effects between internal representations.

4.3.3 Simulation III: Rule-Plus-Exception Category Learning

4.3.3.1 Rationale

In Chapter 3, the work of Erickson and Kruschke (1998) was replicated by applying the ALCOVE model to fit the training and test data observed in their Experiment 1. The results showed that, though the ALCOVE model did a fairly good job on fitting the training data, it was unable to predict the rule-based extrapolation observed in the test phase of the experiment. This was argued to be because ALCOVE was not developed to account for multiple representations in category learning. ALCOVE only emphasises the difference on dimensional attention strengths. In contrast, as shown in previous reimplementation and simulations, SUSTAIN can not only embody attention learning but also establish different forms of category representations via the

substructure recruitment mechanism. More importantly, Section 4.3.2 has shown that SUSTAIN is able to form fairly flexible category representations. According to these features of SUSTAIN, a natural question to ask, thus, is whether the SUSTAIN model is able to account for heterogeneous category representations? This third simulation study is intended to answer this question, by using SUSTAIN to simulate the experiment of Erickson and Kruschke (1998).

4.3.3.2 Method

The simulation used the randomised trial-by-trial stimuli analogous to those like participants saw in the experiment of Erickson and Kruschke (1998). As shown in Chapter 3, the stimuli varied along two dimensions: height of rectangle (y-axis) and line segment position (x-axis). A schematic representation of the category structure is shown in Fig 3-3.

The model was trained over the course of 29 blocks of 14 trials (i.e., 10 rule instances presented once and 2 exception instances presented twice during each block). After the training trials concluded, the model was tested by letting it assign probability of category membership to each test stimuli.

All four parameters were considered for model fitting. The best fitting values were found by using the Monte Carlo (MC) method. 500 MC searching steps (random sampling in the parameter space) were included. For each step, 100 different training sequences were simulated (i.e., 100 simulated learners). Sum-squared deviation (SSD) between model predictions averaged across simulated learners and target category assignments for each stimulus type was used for parameter estimation.

4.3.3.3 Results

The best fitting values of the parameters are shown in Table 4-3. SUSTAIN's performance with these values is shown in Fig 4-8. As can be seen in the figure, the SUSTAIN model did a good job in fitting the training data, but it performed a little faster than human participants in learning the exceptions. For the test set, though, Fig 4-8 shows that SUSTAIN did not form exemplar-based similarity representations, like ALCOVE. The predictions on two critical items, near rule item T_R ($P = 24.73\%$) and near exception item T_E ($P = 24.73\%$), are substantially different from human data ($P(T_R) = 10.00\%$ and $P(T_E) = 11.00\%$). It seems that SUSTAIN made

category responses on critical items near chance. In addition, the learned dimensional attention strengths and the histogram substructures recruitment (i.e., around 75% of the simulated learners recruited more than 12 substructures) revealed that SUSTAIN learned to use the II strategy, but not rule-based extrapolation (see Fig 4-9).

Table 4-3.

Best fitting values of parameters for simulation III.

| Parameter | Best Fitting Values |
|-----------|---------------------|
| ϕ | 8.0141 |
| r | 13.8900 |
| β | 0.6494 |
| λ | 0.1127 |

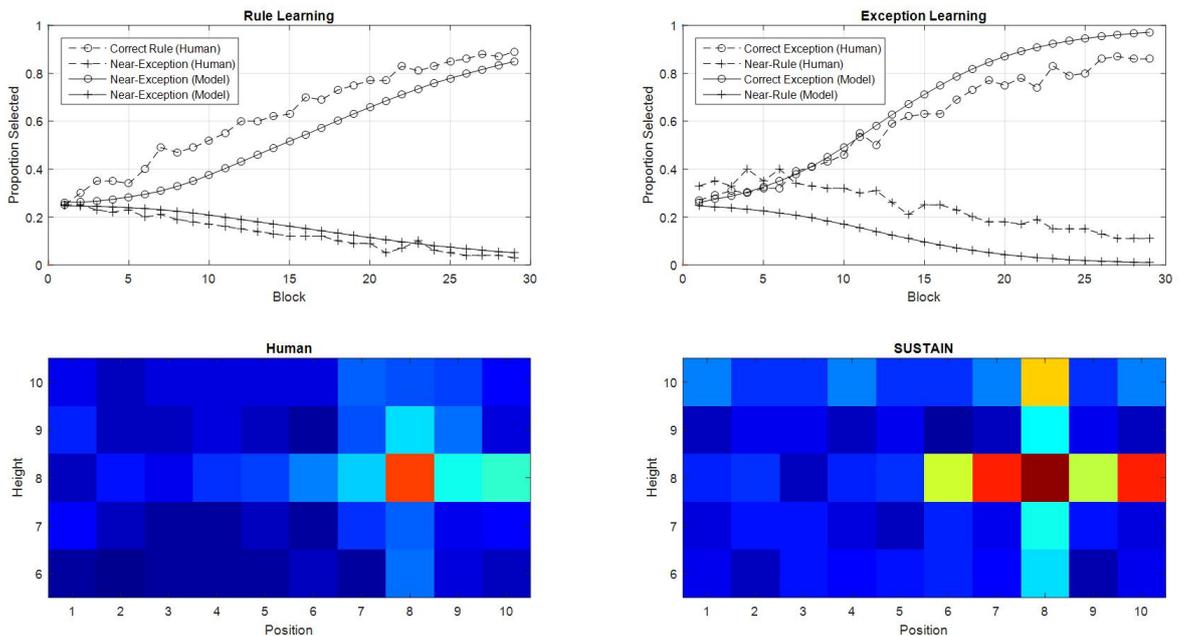


Fig 4-8. Top: The SUSTAIN’s prediction of proportion of rule and exception responses as a function of learning blocks. Bottom: The proportion of exception responses in the test phase observed human participants’ data in Erickson and Kruschke (1998) Experiment 1 and SUSTAIN’S prediction. Cool colours represent low choice probability whereas warm colours represent high choice probability.

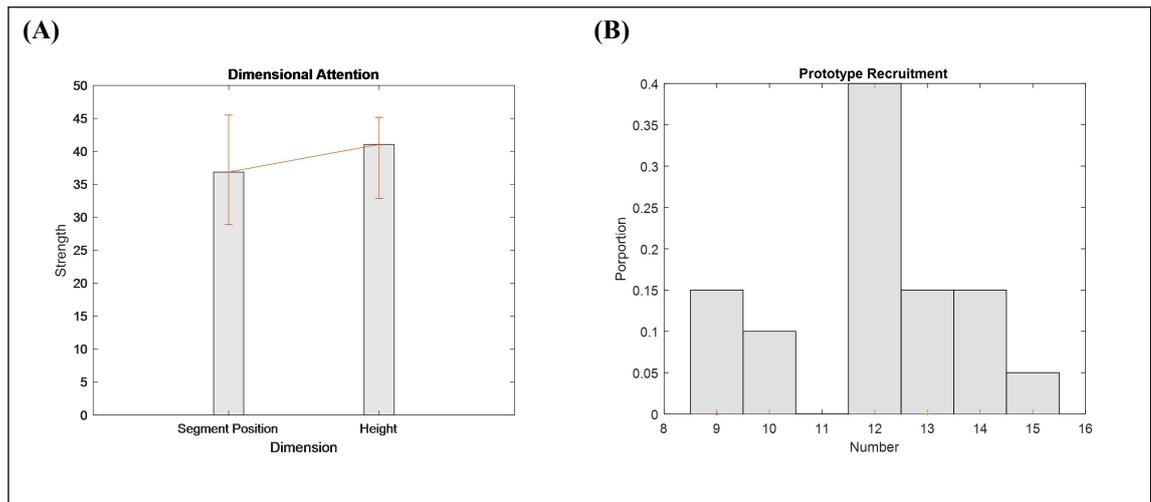


Fig 4-9. (A) Learned attention strengths on each dimension for rule-plus-exception category structures (error bars indicate the maximum and minimum values), and (B) Histogram of distribution of probabilities the model recruiting certain number of prototypes as a result of learning.

4.3.3.4 Discussion

Erickson and Kruschke (1998)'s rule-plus-exception experiment is one of the classic category learning paradigms. This rule-plus-exception representation remains challenging for exemplar-based and prototype-based machine learning models. For example, Nosofsky and Johansen (2000) argued that GCM can properly perform this experiment. However, further follow-up research showed that though the exemplar-based models might fit one experimental dataset, they failed to account for other more complex datasets (Denton et al., 2008; Erickson & Kruschke, 2002; Rodrigues & Murre, 2007). This simulation indicates that the cluster-based SUSTAIN model may not be a solution for rule-plus-exception representation either.

4.3.4 Summary

In this section, the supervised learning version of the SUSTAIN network has been applied to simulate learning of continuous-valued category structures with bivariate and uniform normal distributions, as well as heterogeneous, rule-plus-exception category learning. In these simulations, on the one hand, the model successfully predicted the dissociation between multiple category representations. SUSTAIN recruited fewer substructures for RB categories than for II categories. However, on the other hand, SUSTAIN did not predict the inhibition between distinct

representations in hybrid category learning. Although the model produced greater strength on the bar frequency dimension than on the orientation dimension, as it was unable to neglect the influence from the orientation dimension, it seems that SUSTAIN successfully learned the hybrid decision boundary which is inconsistent with results from Ashby and Crossley (2010). Moreover, SUSTAIN failed to predict the rule-based extrapolation effects. The probabilities of exception responses for two critical test items were near chance, which is inconsistent with the human data.

4.4 Theoretical Implications

4.4.1 Limitations of SUSTAIN

SUSTAIN suggests a multiple-competing-internal-representations approach, which postulates that category learning begins with simple solutions, but more and more substructures are recruited as the difficulty of the task increases. As can be seen in the above simulations, more prototypes are needed for learning complex tasks, such as Shepard task Types III, IV and V and II categories, than for learning simple categories, such as Shepard task Types I and II. As more and more substructures are recruited, learning in SUSTAIN finally yields an exemplar-like process. This approach is suitable for learning binary-valued category structures (Love et al., 2004).

However, according to its performance on the simulations presented here, SUSTAIN has three limitations. First, SUSTAIN's attention learning mechanism cannot ignore the influence from irrelevant dimension. As mentioned earlier, each dimensional attention strength cannot be less than 1.0. Second, the multiple-competing-internal-representations approach of SUSTAIN is too flexible to reflect the nature of multiple representations in category learning. Third, SUSTAIN partly inherited ALCOVE's attention learning framework, suggesting that dissociations between representations results from the formed substructures of tasks, but not differences on dimensional attention distribution between internal representations. SUSTAIN is not able to implement attentional shifts in a heterogeneous category learning task. These

shortcoming prevent SUSTAIN from becoming a candidate for developing an integrated account of category learning.

4.4.2 The Organisation of Internal Representations

Love et al. (2004) argued that the category representation a human learner forms should be highly dependent on the current goals of the learner (see also Love, 2003; Ross, 1996; 1997). The internal representations of categories are organised to serve these goals. An integrated account of category learning should be able to address a range of goals, tasks and functions. SUSTAIN was intended as an account of how humans incrementally discover the internal representations of categories based on goals. Although, the implementation of SUSTAIN failed to account for multiple representations in the continuous-valued category structure, the emphasis of the role of goals in category learning is meaningful for theoretical integration. The goal representations should be considered in the organisation of internal representations of category learning.

4.4.3 Relationship to Modular Architecture

In addition to SUSTAIN, there are two modular network models based on multiple representations of category learning, COVIS and ATRIUM. The modular networks suggest that there are two types of category representation systems: a rule-based representation system and an exemplar-based representation system. Incorporating these competing systems requires inclusion of mechanisms associated with resolving this competition and creates a potential credit assignment problem in handling feedback. For doing so, COVIS postulates an overall system weighting mechanism determined by the confidence and history of accuracy for each subsystem, whereas ATRIUM applies a gating network. The overall system weighting mechanism successfully predicts that the rule-based subsystem dominates learning hybrid categories, while exemplar-based subsystem dominates II learning. As mentioned earlier, both SUSTAIN and the modular networks are influenced by exemplar theory. ATRIUM extends ALCOVE by

introducing the extra rule representation and the modular architecture, whereas COVIS retains the exemplar network to represent a procedural learning system. SUSTAIN introduces a more economical approach in its storage of examples – it only stores exemplars as separate internal representations when a prediction error occurs. The SUSTAIN approach is also inspired by a symbolic rule-plus-exception (RULEX) model in which simple rules (e.g., one dimensional rules) are considered first, and exceptions are encoded, or more complex rules are considered, when the categories are not mastered by the rule.

Although the multiple-prototypes network and the modular networks are architecturally different, all of them share the idea that there are separate competing internal representations in the human category learning system. How does the system resolve the competition between internal representations? This is the problem where establishing a control mechanism is needed. SUSTAIN and ATRIUM suggest a flexible, item-specific control mechanism, whereas COVIS suggests a global control mechanism. However, evidence from empirical studies of category learning indicates that both mechanisms exist. In a sense, a complete account of control of the organisation of internal representations should be able to use different strategies in response to the different types of categories encountered within the environment.

Chapter 5.

Case Study II: Multiple Systems Theory and Modular Architecture of COVIS

5.1 Introduction

Although the SUSTAIN model (Love et al., 2004) embodies multiple representations in category learning, it implements the traditional feed-forward network architecture, arguing that the formation of distinct category representations is determined by attention learning and recruitment of surprising events. This network architecture can account for the dissociation between RB and II category learning, but fails to account for effects related to hybrid and heterogeneous category representations. It, thus, motivates us to give insights into other multiple representations accounts of category learning.

In 1998, some researchers argued that the category learning system should have a modular architecture (Ashby et al., 1998; Erickson & Kruschke, 1998). In particular, Ashby et al. (1998) proposed a computational cognitive neuroscience model, the COVIS (COmpetition between Verbal and Implicit Systems) model, arguing that the modular architecture should, at least, include an explicit learning system and an implicit learning system.

Computational cognitive neuroscience (CCN) is an approach that lies at the intersection of cognitive neuroscience and modern machine learning theory (i.e., neural networks). Unlike traditional machine learning approaches, CCN suggests that in addition to predicting human behavioural data, it is important to have neurobiological accuracy in computational modelling. As a result, CCN models commonly have many more neurobiological constraints than traditional

machine learning approaches. CCN researchers suggest that although the modern neural network models have some features in common with the human brain, such as changes in synaptic strengths and continuous flow of activation, there is almost no attempt to identify units in network models with specific brain regions (or neural pathways). In addition, it is argued that the machine learning rules used in most network models do not precisely reflect biological plausibility (e.g., Ashby, 2018; Ashby & Rosedahl, 2017).

According to Marr's (1982) three levels of analysis, mathematical modelling could be divided into computational, algorithmic and implementational. Computational-level models make quantitative predictions, but do not describe the algorithms that produce those prediction. Algorithmic-level models describe the algorithms, but not the architecture that implements these algorithms. Finally, the implementational-level models describe the architecture that implements the algorithms that produce the quantitative predictions. In fact, the focus of most traditional network models is to model complex behaviours at an algorithmic level, because before the cognitive neuroscience revolution of 1990s, understanding of neurobiological basis of behaviour was fairly difficult. As more and more cognitive neuroscience data has been provided, more neurobiologically plausible learning algorithms and models of individual units have been proposed (Ashby et al., 1998; Cohen et al., 1996).

The original version of the computational implementation of COVIS is one of these CCN models, the focus of which is on the implementation of a neurobiologically-constrained architecture. COVIS has achieved great success in predicting both behavioural data (e.g., Ashby & Waldron, 1999; Maddox & Ashby, 1993; Waldron & Ashby, 2001) and neuroscience data (e.g., Ashby & Ell, 2001; 2002; Knowlton et al, 1996; Maddox & Filoteo, 2001). However, no attempt has been made to investigate the control mechanism within this modular architecture that determines response generation based on multiple internal representations. Therefore, this chapter presents a simulation study to demonstrate how COVIS accounts for the control of multiple representations in category learning.

5.2 COVIS

5.2.1 The Neurobiological Theory

COVIS postulates two systems that compete throughout learning: an explicit learning system that learns verbalisable rules and a procedural learning system that uses a form of implicit and incremental learning. This assumption is based on the widely accepted multiple memory systems theory (Squire, 1992). In the multiple memory systems theory, it is assumed that human memory consists of, at least, two parts, namely declarative memory and nondeclarative memory. The declarative memory refers to the memory that provides the basis for conscious and explicit reasoning and recollections of facts and events. In contrast, the nondeclarative memory system seems not to need the participation of consciousness. The idea of nondeclarative memory stems from research with amnesic patients. Amnesic patients commonly perform poorly on declarative memory tasks (e.g., recall and recognition), but they often show relatively intact memory on nondeclarative memory tests, such as repetition priming on word stem completion tasks (Graf & Schacter, 1985; Schacter & Buckner, 1998; Schacter & Graf, 1986). In a study by Knowlton et al. (1996), an amnesic group demonstrated comparable performance to a control group on a classification task—the weather prediction task (WPT), but their performance was disproportionately impaired on a recognition task. Neurobiologically, it was argued that the acquisition of declarative memory is associated with prefrontal cortex (PFC), while acquisition of nondeclarative memory depends on the basal ganglia (e.g., Daw et al., 2005; Otto et al., 2014; Seger & Spiering, 2011; Schultz et al., 1997).

In the past decade, a large and growing body of research has suggested that humans have available multiple processing modes that can be used during categorisation. Considerable evidence has shown that the declarative and nondeclarative memory systems both contribute to perceptual category learning. In particular, Ashby et al. (1998) proposed in COVIS that categorisation is mediated by two cortico-striatal circuits acting in parallel, a prefrontal-based explicit learning system (see Fig 5-1A) and a striatal-based procedural learning system (see Fig 5-1B). The operation of the explicit learning system is flexible, while the procedural learning

system is implicit, inflexible and automatic. As mentioned in Chapter 2, much evidence of COVIS comes from RB and II category learning tasks (see Section 2.2 for more detail).

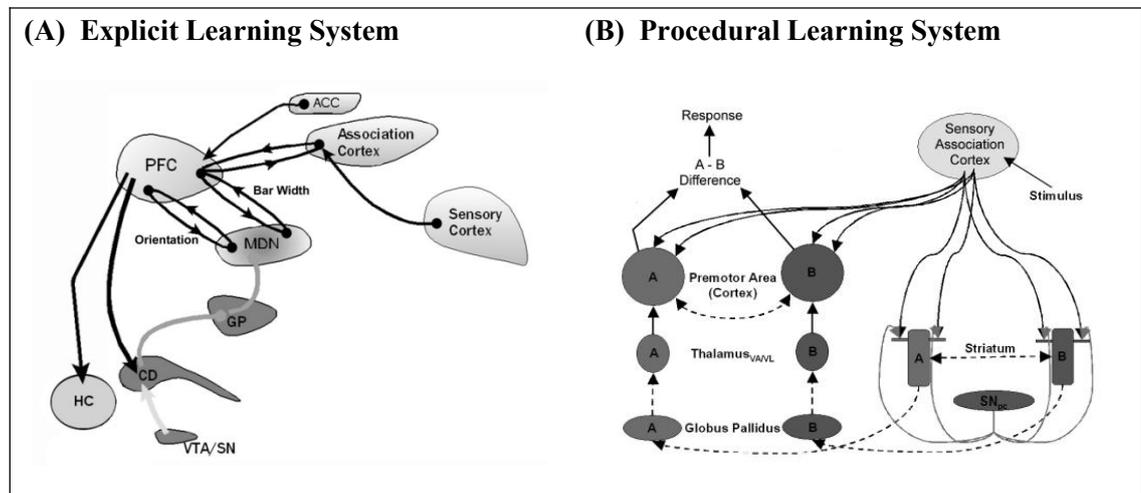


Fig 5-1. Illustration of separate systems of COVIS account. **Left panel:** hypothesis testing system (black arrows = excitatory projections, gray arrows = inhibitory projections, light gray arrow = dopaminergic projections, VTA = ventral tegmental area, SN = substantia nigra pars compacta, CD = caudate nucleus, GP = globus pallidus, MDN = medial dorsal nucleus (of the thalamus), PFC = prefrontal cortex, ACC = anterior cingulate cortex, HC = hippocampus); **right panel:** procedural learning system (Excitatory projections end in solid circles, inhibitory projections end in open circles, and dopaminergic projection is dashed. PFC = prefrontal cortex, Cau = caudate nucleus, GP = globus pallidus, and Th = thalamus).

As can be seen in Fig 5-1A, the key structures of the explicit learning system are the prefrontal cortex (PFC), anterior cingulate cortex (ACC), caudate nucleus and hippocampus. There are two subsystems: a maintenance network that maintains rules in working memory, tests and mediates switching between candidate rules, and a selection network that produces and selects candidate rules. The maintenance network includes all structures in Fig 5-1A except the ACC. Two reverberating cortical-thalamic loops maintain activation in the PFC working memory units during explicit learning. The cortical input units send excitatory signals to lateral PFC (working memory area), which can send excitatory signals back to the same cortical units, thereby forming the first reverberating loop. The second reverberating loop consists of projections between PFC and the medial dorsal (MD) nucleus of the thalamus (Alexander et al, 1986). However, the high spontaneous activity that is characteristic of the GABAergic neurons in the globus pallidus (GPi) tonically inhibit the thalamus, which prevents the closing of this cortical-thalamic loop, leading to the loss of information from working memory. To deal with this inhibition, the PFC may excite the medium spiny neurons (MSNs) in the head of the caudate

nucleus (Bennett & Wilson, 2000), which in turn inhibit the GABAergic neurons that are inhibiting the thalamus. Reducing the pallidal inhibition of the thalamus allows reverberation in cortical-thalamic loops, and thereby facilitates operation of the maintenance network.

The operations of the selection network are mediated by separate neural processes. The production of new rules is complex, and a complete model has not yet been formulated. In the original implementation of COVIS, the ACC selects among alternative rules by enhancing the activity of the specific lateral PFC unit that represents a particular rule (Ashby, Valentin, & Turken, 2002). COVIS predicts that the most effortful and time-consuming processing follows error feedback, because negative feedback suggests that the current rule is incorrect. In support of this prediction, during the process of initial learning, slower response times on trials following an error predict higher accuracy in RB tasks, but not in II tasks (Tam, Maddox, & Huang-Pollock, 2013). When this process leads to the selection of the correct rule, there is often a sudden transition from suboptimal to optimal performance accompanied by an abrupt transition in neural firing of PFC neuron ensembles (Durstewitz, Vittoz, Floresco, & Seamans, 2010; Smith & Ell, 2015).

In contrast, the key structure of the procedural learning system is the striatum, the major input region of basal ganglia (BG) that consists of the caudate nucleus and the putamen. As the procedural learning system reflects the incremental stimulus-response associations via dopamine-mediated Hebbian learning, it is assumed that all extrastriate visual units project directly to the striatum, with about 10,000 visual neurons converging on such striatal MSNs (Wilson, 1995). Through learning, each MSN associates an abstract motor program with a number of visual neurons. The striatal MSNs receive input from a variety of cortical areas, including PFC and premotor areas and send their axons to BG output structures (Bolam et al., 1985). There exist two synapses on this pathway. The first synapse is in the internal segment of the GPi, which is the output structure within the BG, and the second synapse is in the thalamus. The dorsal putamen projects primarily into premotor areas of cortex (e.g., supplementary motor area, SMA) via the ventral lateral nucleus of the thalamus (Matelli & Luppino, 1996). The SMA is interconnected with the premotor areas (Dum & Strick, 2005). In contrast, the caudate and anterior putamen project to cortex via the MD and ventral anterior (VA) thalamic nuclei. The MD nucleus projects into anterior areas of frontal cortex, including PFC, whereas the primary

cortical projection from VA is to preSMA (Matelli & Luppino, 1996), which is interconnected with the PFC (Akkal et al., 2007; Wang et al., 2005).

Learning in the procedural-learning system is determined by the DA-mediated reward signals from the substantia nigra pars compacta (SNpc). It is argued that a primary function of DA is to serve as the reward signal in reinforcement learning (e.g., Houk et al., 1995; Wickens, 1993). The positive feedback that follows successful behaviours increases phasic DA levels in the striatum, which strengthens recently active synapses, whereas negative feedback causes DA levels to fall below baseline, which weakens recently active synapses (e.g., Arbuthnott, Ingham, & Wickens, 2000; Calabresi, Pisani, Mercuri, & Bernardi, 1996; Reynolds & Wickens, 2002). In other words, the DA levels serve as a teaching signal for which successful behaviours increase in probability and unsuccessful behaviours decrease in probability. In this sense, synaptic plasticity, including long-term potentiation (LTP) and long-term depression (LTD), can only occur when the visual trace of the stimulus and the postsynaptic effects of DA overlap in time.

5.2.2 Model Description

Explicit Learning System The explicit-learning system selects and tests verbal rules that determine category membership. The computational implementation of the explicit learning system is a hybrid network that includes both symbolic rule selection with a switching component and a connectionist salience learning component. In most applications, the model initially contains a set of explicit rules (e.g., one-dimensional rules, conjunctive rules, or disjunctive rules). The initial salience of each rule is set according to certain principles. It is assumed that rules that people have abundant prior experience with have high initial salience, and rules that have rarely been used before have low initial salience. In typical applications, the initial saliences of all one-dimensional rules are set equal, whereas the initial saliences of conjunctive and disjunctive rules are set much lower.

In the explicit-learning system, a response is selected by computing a discriminant value $h_E(x)$, defined as:

$$h_E(x) = x_i - C_i - \varepsilon_E \quad (5.1)$$

where x_i is the value of stimulus x on dimension i , C_i is a decision criterion and ε_E is random noise (a normally distributed random variable with mean 0 and variance α_E^2). The response rule is: respond A on trial n if $h_E(x) < 0$, otherwise, respond B.

The salience of each rule is adjusted after every trial on which it is used in a manner that depends on whether or not it is correct. Let $Z_K(n)$ denote the salience of rule K on trial n . The learning rule is straightforwardly expressed as:

$$Z_K(n) = \begin{cases} Z_K(n-1) + \Delta_C & \text{if correct} \\ Z_K(n-1) - \Delta_E & \text{if incorrect} \end{cases} \quad (5.2)$$

where Δ_C and Δ_E are positive constants. Δ_C represents the gain associated with a correct response and Δ_E represents the cost associated with error. Selection and switching among all rules is finally determined by the selection weight of each rule, Y , adjusted from each rule's salience. It is expressed as

$$\begin{cases} Y_i(n) = Z_i(n) + \gamma, & \text{if rule } i \text{ is active on trial } n \\ Y_j(n) = Z_j(n) + X & \text{if rule } j \text{ is randomly selected on trial } n \\ Y_k(n) = Z_k(n) & \text{otherwise} \end{cases} \quad (5.3)$$

where γ is a constant called the perseveration parameter and X is a random variable that has a Poisson distribution with mean λ . λ is a constant called the selection parameter. Selection and switching among all rules is determined by each rule's weight.

The explicit-learning system of COVIS has shown some advantages in accounting for rule-based or rule-like categorisation tasks (Helie et al., 2011). In particular, the involvement of perseveration and selection processes may facilitate its success in simulating behaviour on rule switching tasks, such as the Wisconsin Card Sorting Task (e.g., Helie et al., 2012a).

Procedural-Learning System The original version of procedural learning system is a two-layer network model. It is assumed that the key site of learning is at cortical–striatal synapses within the striatum. Therefore, the procedural-learning system is also referred to as the

striatal pattern classifier (SPC, Ashby & Waldron, 1999). The implementation of SPC is fairly straightforward. The activation in input units is given by

$$a_k^{in} = \exp\left(\frac{-d(K, stimulus)^2}{\sigma}\right) \quad (5.4)$$

where d is the Euclidean distance between stimulus and input unit K , and σ governs the width of the radial basis function. The output layer in the procedural-learning system is assumed to represent the striatum. The activation in each striatal unit on trial n , $S_j(n)$, is determined by the weighted sum of activations in all input units. The response rule is: respond J , if $S_j(n) = \max(S_j(n))$.

The weights in SPC are updated based on the three factors: (1) pre-synaptic activation, (2) post-synaptic activation, and (3) dopamine levels. The learning rule is expressed as:

$$w_{j,k}(n+1) = w_{j,k}(n) + \begin{cases} \alpha_w (S_j(n) - \theta_{MND A})(DA(n) - DA_{base})(1 - w_{j,k}(n))a_k^{in} & \text{if } S_j(n) > \theta_{MND A} \text{ and } DA(n) > DA_{base} \\ \beta_w (S_j(n) - \theta_{MND A})(DA(n) - DA_{base})w_{j,k}(n)a_k^{in} & \text{if } S_j(n) > \theta_{MND A} \text{ and } DA(n) < DA_{base} \\ \gamma_w (S_j(n) - \theta_{MND A})(\theta_{AMPA} - S_j(n))w_{j,k}(n)a_k^{in} & \text{if } S_j(n) < \theta_{MND A} \text{ and } S_j(n) > \theta_{AMPA} \end{cases} \quad (5.5)$$

where DA_{base} is the baseline dopamine level and $DA(n)$ is the dopamine released on trial n following feedback (see below). α_w , β_w and γ_w are learning rates which determine the magnitudes of increases and decreases in synaptic strength. θ_{NMDA} and θ_{AMPA} represent the activation thresholds for post-synaptic NMDA and AMPA glutamate receptors, respectively. The numerical value of $\theta_{NMDA} > \theta_{AMPA}$ because NMDA receptors have a higher threshold for activation than AMPA receptors. This is critical because NMDA receptor activation is required to strengthen cortical-striatal synapses (Calabresi et al., 1996).

Dopamine Model Changes of associative weights between sensory units and striatal units are determined by the amount of dopamine released on every trial in response to the feedback signal. The procedural system of COVIS uses a simple model of dopamine release by first computing both obtained (R_n) and predicted reward (P_n), and then by estimating the amount

of dopamine released as a function of reward prediction error (RPE), where RPE = obtained reward – predicted reward. The obtained reward on trial n is defined as:

$$R_n = \begin{cases} +1, & \text{if feedback is positive} \\ 0, & \text{if no feedback} \\ -1, & \text{if feedback is negative} \end{cases} \quad (5.6)$$

The predicted reward is defined as:

$$P_n = P_{n-1} + \alpha_{pr}[R_{n-1} - P_{n-1}] \quad (5.7)$$

where α_{pr} governs how quickly P_n converges. To compute the dopamine release on each trial, a simple model matching empirical results reported by Bayer and Glimcher (2005) is used:

$$DA(n) = \begin{cases} 1 & \text{if RPE} > 1 \\ .8RPE + .2 & \text{if } -.25 < \text{RPE} < 1 \\ 0 & \text{if RPE} < -.25 \end{cases} \quad (5.8)$$

Note that it is assumed that, for modelling normal adults, the baseline dopamine level is set to .2 and that dopamine levels increase linearly with RPE between a floor of 0 and a ceiling of 1.0.

Resolving System Competition In order to resolve the competition between the systems and to select an overall response, confidence and trust of each system are measured on every trial. System confidence is calculated as a discriminant value for each system. In a case with two alternative categories, the discriminant value for the procedural-learning system $h_p = |S_A(n) - S_B(n)|$. The trust placed in each system is determined by overall system weights, θ_E and θ_P , where $\theta_E + \theta_P = 1$. In a typical application, θ_E is often set to .99. The system weights are updated based on the success of the explicit learning system, which is defined as:

$$\theta_E(n+1) = \begin{cases} \theta_E(n) + \Delta_{OC}(1 - \theta_E(n)) & \text{if explicit learning system is correct} \\ \theta_E(n) - \Delta_{OE}\theta_E(n) & \text{otherwise} \end{cases} \quad (5.9)$$

where Δ_{OC} and Δ_{OE} are learning rates. Finally, the overall system decision rule is to emit the response suggested by the explicit system if $\theta_E \times |h_E| > \theta_P \times |h_P|$; otherwise, it emits the response suggested by the procedural system.

5.2.3 Control Mechanisms in COVIS

To summarise, COVIS, as a modular architecture, has two types of control mechanisms in the category learning system. One control mechanism is cognitive control in the explicit learning system. As the explicit learning system embodies competition between rule-based representations acting in parallel, the structure of the system, too, follows a modular architecture. COVIS suggests that switching and selection between verbal rules is related to cognitive control. Each rule's salience changes on trial and error, whereas the weights for selecting the rules are determined by the switching parameter λ and the perseveration parameter γ . Based on this mechanism, the model can switch away from rules that produce errors and persevere with the rules that are more rewarded than others. This mechanism is suitable for fast RB category learning.

The second control mechanism is to use the confidence and trust of each system to resolve system competition. Because Ashby et al. (1998, 2011) argued that procedural learning system is independent of cognitive control, the cognitive control mechanism is rejected as resolving the competition between multiple systems. The procedural learning system globally wins the competition once its overall performance is better than the explicit learning system. This mechanism suites II category learning. But, is it sufficient to reflect the interaction between multiple systems?

For a homogeneous category learning task, switching of control on a trial-by-trial basis is not required, whereas, for a heterogeneous category learning task, task partitioning and trial-by-trial switching is essential. Kruschke and his colleagues argued that, as the competition between multiple systems in COVIS is driven by a separate heuristic that combines the long-term accuracies of the modules with the decisiveness of each module regarding the current stimulus, the control mechanism in COVIS cannot yield the necessary control passed back and forth between the systems on a trial-by-trial basis (e.g., Erickson & Kruschke, 1998; Kruschke, 2011). However, Ashby and his colleagues argued that COVIS can implement system switching on a trial-by-trial basis (e.g., Ashby & Crossley, 2010; Ashby et al., 1998; Crossley et al., 2017; Paul & Ashby, 2013). These two arguments have caused great confusion in the literature. To

demonstrate the capacity of this control mechanism, the next section presents a simulation that demonstrates how COVIS accounts for hybrid category learning.

5.3 Simulation Study: Hybrid Category Learning

5.3.1 Rationale

Human participants reliably learn II categories. COVIS was initially developed for this basic performance benchmark. COVIS postulates that, over the course of training, human participants identify which system is best suited for the current task. The system comes to dominate performance and the model essentially reduces to a single system. An empirical study of human hybrid category learning supports this assumption. In a hybrid category learning task a 1D rule-based strategy is required on some trials and an II strategy is required on other trials to achieve maximum accuracy, but participants have difficulty switching between strategies and instead tend to use the 1D rule-based strategy for all trials.

In this simulation, the purpose is not to show how well the model can fit some data set, but to illustrate how COVIS learns and how control of responding is transferred from the explicit learning system to the procedural learning system in different category structures. The schematic representations of category structures used are shown in Fig 5-2. For the II categories, a total of 300 stimuli were drawn from each of two bivariate normal distributions. Category means and variances were identical to that of Paul and Ashby (2013) (see Table 5-1). The suboptimal one-dimensional rule can only reach 78.5% accuracy. For the hybrid categories, a total of 300 stimuli were randomly sampled from each of two categories separated by a hybrid boundary (Fig 5-2B). The maximum performance by the suboptimal one-dimensional rule on these categories is about 90%.

Table 5-1.
Parameters (i.e., mean, variance and covariance) II categories used in simulation.

| Category | μ_X | μ_Y | σ_X^2 | σ_Y^2 | Cov _{X,Y} |
|----------|---------|---------|--------------|--------------|--------------------|
| A | 40 | 60 | 167.59 | 167.59 | 151.26 |
| B | 60 | 40 | 167.59 | 167.59 | 151.26 |

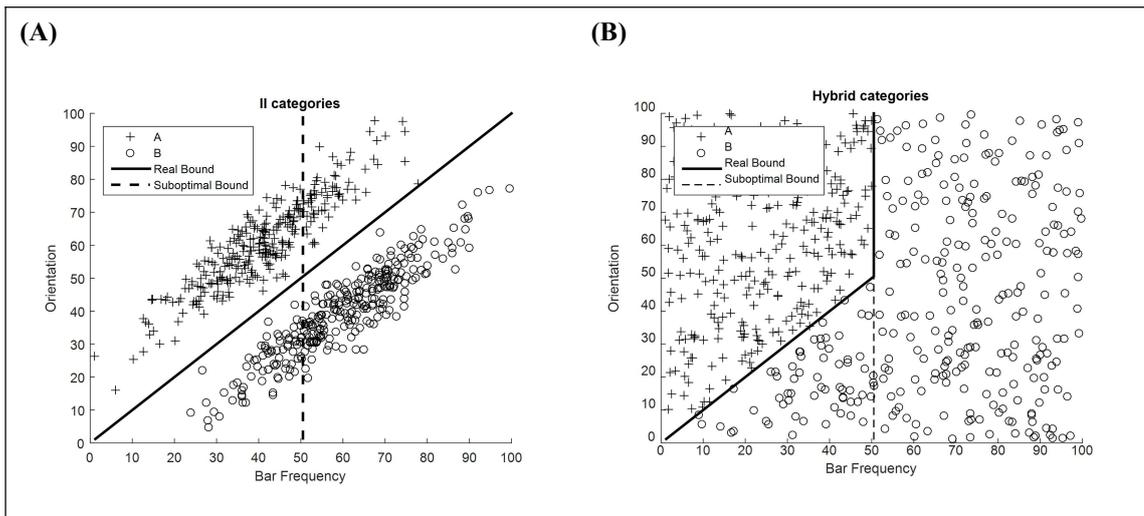


Fig 5-2. Schematic representation of information-integration (II) category structure used in the simulation. Dot marks (asterisks and circles) represent the 2D coordinates of stimuli (e.g., disks varying in spatial frequency and orientation). The solid black line represents the optimal boundary, the dashed block line represents a suboptimal rule-based boundary.

5.3.2 Implementation Details

The learning mechanism of the COVIS model is quite straightforward. However, in the literature, the various implementations of COVIS differ (e.g., Ashby et al., 1998; 2011; Helie et al., 2012a; 2012b; Paul & Ashby, 2013). For example, in Ashby et al. (2011) separate systems receive feedback signals independently whereas the systems receive a single feedback signal in an alternative version (Paul & Ashby, 2013).

The simulated reported here implements a simplified version of COVIS which has proven to be helpful for illustrative purposes (Ashby et al., 2011; Paul & Ashby, 2013). The implementation is simplified in three ways. First, random noise variables at the decision level of both systems were ignored. Thus, the discriminant values of each system are deterministic. Second, a response in the explicit learning system is produced by the most likely suboptimal one-dimension rule from the outset (i.e., no rule selection and testing is manipulated here). Third, in the procedural learning system, a strong form of lateral inhibition between striatal units is assumed. In other words, only the weights associated with the most activated striatal units are updated on each trial. These simplifications improve the computational efficiency of the model.

5.3.3 Method

Learning in COVIS could be influenced by several factors. First, COVIS is sensitive to the randomised stimulus ordering and initialised random weights in the procedural learning system. To handle the effects of these random variables, the results reported below are based on a fixed random sample set of 15 random stimulus orderings and weight initialisations. Each simulation in the parameter search stage is tested on this set, and model performance is averaged across them. Second, performance of the COVIS model could be influenced by several parameters (see above). However, including every parameter of the model would be inefficient because the parameter space would be very high dimensional and many parameters interact in predictable ways. Because the purpose is to illustrate COVIS's ability of learning and system switching, the simulation considers the effects of three learning rate parameters in the procedural learning system (α_w , β_w and γ_w) and two system switching parameters (Δ_{OC} and Δ_{OE}). These five parameters directly affect how fast the procedural learning system can learn and how fast control is transferred. Table 5-2 summarises the function and search range for each selected parameter. Every other parameter is fixed to a specific value ($\theta_{AMPA} = 0.01$ and $\theta_{NMDA} = 0.1$).

Table 5-2.
Function and search range for each parameter used in the simulation study

| Parameter | Function | Range |
|------------------|--|--------------|
| α_w | Learning rates for strengthening striatal weights | [0.01, 1] |
| β_w | Learning rates for weakening striatal weights (Condition 1) | [0.01, 1] |
| γ_w | Learning rates for weakening striatal weights (Condition 2) | [0.01, 1] |
| Δ_{OC} | Controls how explicit system bias grows for correct responses | [0.001, 0.2] |
| Δ_{OE} | Controls how explicit system bias decays for incorrect responses | [0.001, 0.2] |

The next thing is to decide how to define the measure of the model performance. Here, the mean accuracy of the last 100 trials is used as a measure of learning, because in a typical empirical studies of II category learning, the performance on the last 100 trials of 600 trials is

often used as test trials (e.g., Ashby et al., 2003; Ashby & O'Brien, 2007; Zethamova & Maddox, 2006).

The simulation was run on 10,000 steps. 10,000 points were randomly sampled from the parameter space by the Monte Carlo method. Each step consisted of 15 blocks of II and hybrid category learning including 600 trials with re-initialisation of the model occurring between blocks.

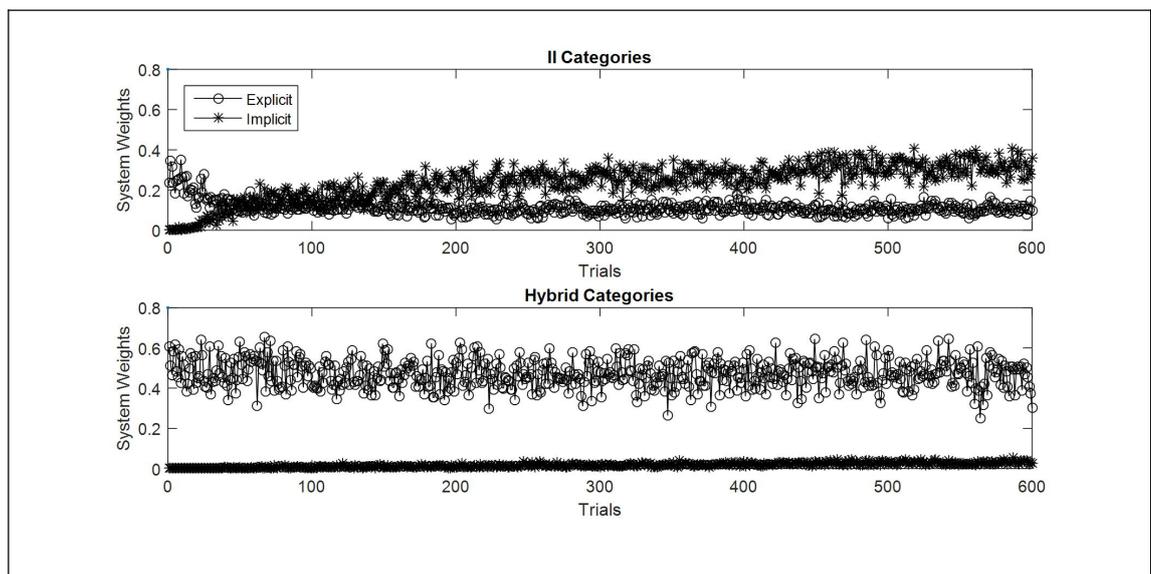


Fig 5-3. Overall system weights change during learning of II (Top) and hybrid (Bottom) categories.

5.3.4 Results and Discussion

Figure 5-3(1) and 5-3(2) show that the procedural learning system learns independently in both tasks. This is partially because the simplified version of COVIS here includes strong lateral inhibition between response units, and this manipulation may facilitate procedural learning.

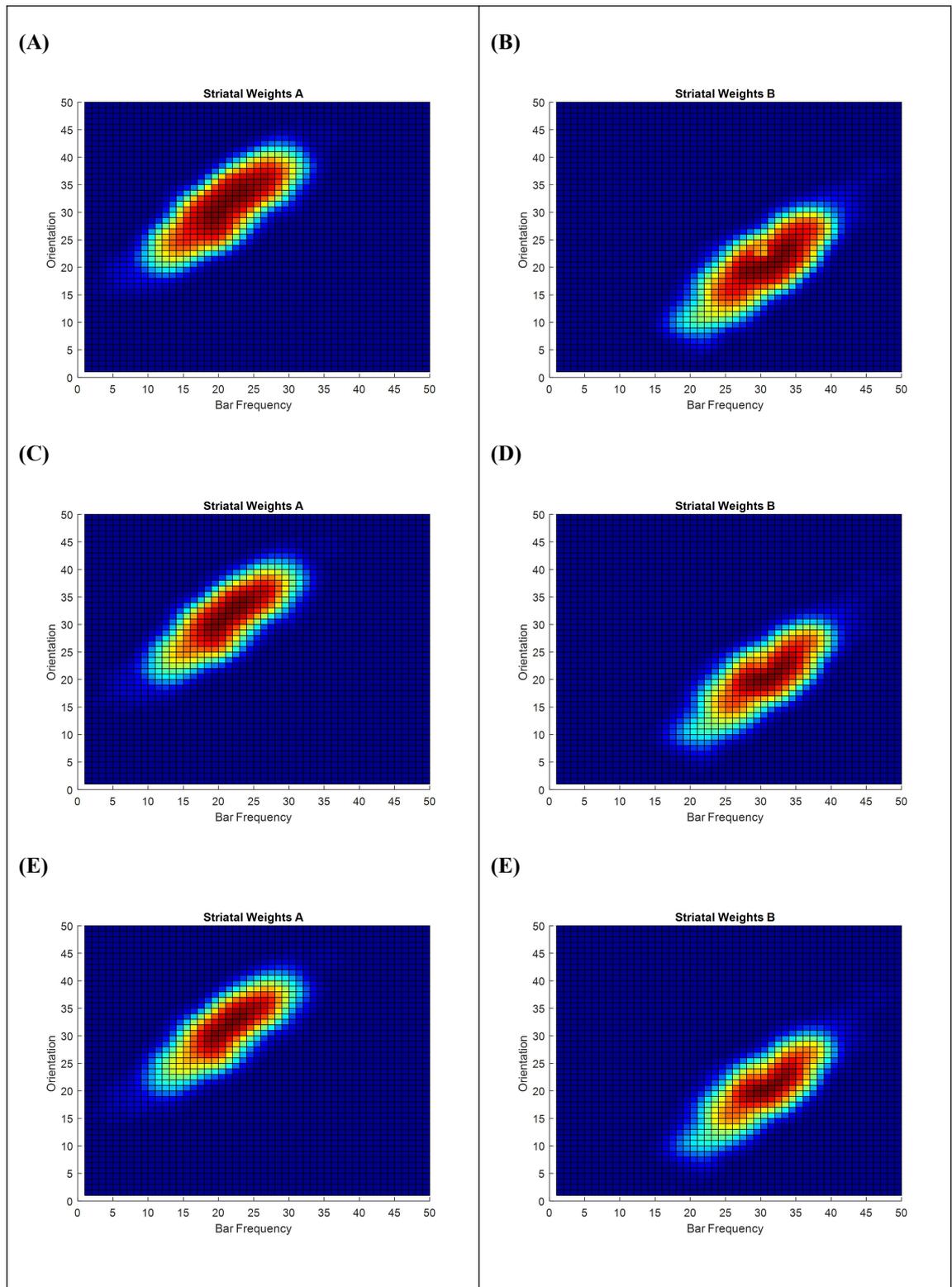


Fig 5-3(1). Striatum weights in the procedural learning system after training in II category structures. The weights are averaged across the output produced by the parameters that produce (Top) performance above 90% accuracy, (Middle) performance ranged between 80% and 90% accuracy; (Bottom) performance ranged between 70% and 80 % accuracy. Cool colours represent small weights whereas warm colours represent large weights.

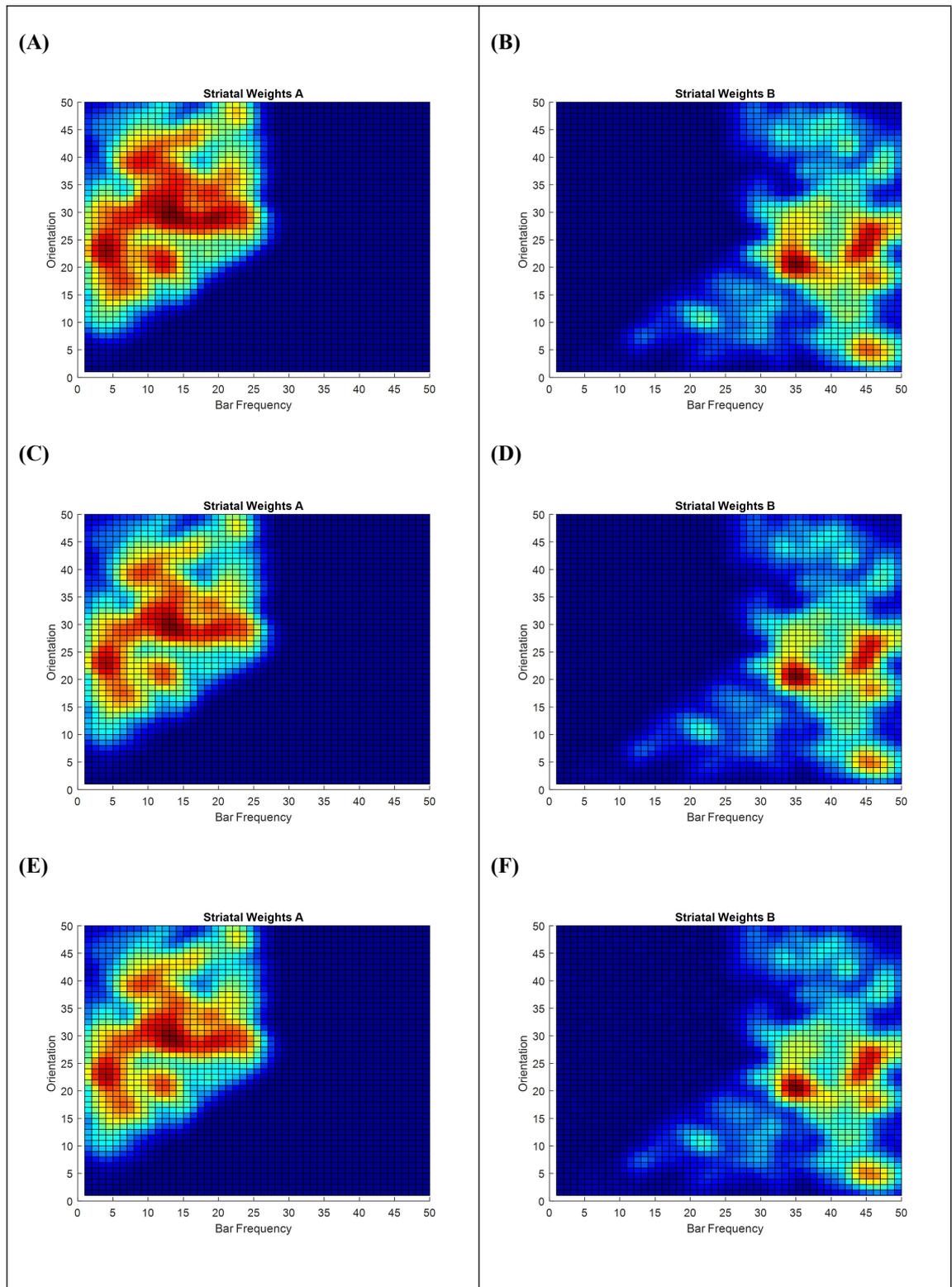


Fig 5-3(2). Striatal weights in the procedural learning system after training in hybrid category structures. The weights are averaged across the output produced by the parameters that produce (Top) performance above 90% accuracy, (Middle) performance ranged between 80% and 90% accuracy; (Bottom) performance ranged between 70% and 80 % accuracy. Cool colours represent small weights whereas warm colours represent large weights.

Fig 5-4 shows how the model learn each category structure. The systems weights are averaged across performance above 90% accuracy (i.e., in order to ensure that the model has learned the tasks). According to Ashby and Crossley (2010), in COVIS, the most confident system controls responses. Ideally, for stimuli with steep orientations, the two systems would have similar levels of confidence, but the model is biased towards the explicit-learning system, so presumably the explicit learning system would control responding for these stimuli. For stimuli with shallow orientations, the explicit learning system would have low confidence, so responding would be controlled by the procedural learning system.

However, Fig 5-4 shows a different patterns of model performance. For II category learning, after over one hundred trials the control of response generation is determined by the procedural learning system. But, for the hybrid category learning, though the procedural learning system learns very well, it appears to be unable to control the response generation. The system switching mechanism in COVIS is, thus, successful in predicting the qualitative difference between learning II and hybrid categories (Ashby and Crossley, 2010). However, a problem may be that COVIS does not perform trial-by-trial switching. Instead, at the end of learning, the model reduces to a single system. At the end of training in the II task, the control of responses is transferred from the explicit learning system to the procedural learning system. Whereas, it appears that the control of responses in the hybrid category learning task does not change during learning.

To summarise, in contrast with the arguments of Ashby and his colleagues, the results of this simulation reveal that COVIS is unable to account for system switching on a trial-by-trial basis. This is indeed to be expected, because the COVIS model was originally designed to account only for the homogeneous category learning tasks, and not for heterogeneous category learning. The module-combination rule for resolving the competition between the systems does not allow system switching on a trial-by-trial basis. Ashby and his colleagues, though, argued that COVIS can be modified to yield task switching (e.g., Paul & Ashby, 2013). It nevertheless remains problematic that COVIS suggests that control of competition between the systems does not involve cognitive control. This issue will be discussed in detail in the next section.

5.4 Theoretical Implications

5.4.1 Limitations of the CCN Approach

As a network model inspired by multiple memory systems theory, the biggest theoretical value of COVIS lies in providing a clear illustration of a dual systems framework. The model's basic assumption is that people begin by searching for verbal rules, if there is no acceptable rule, then control is transferred to the procedural learning system. In other words, the transfer of control is driven by the performance of the explicit learning system. However, recent empirical studies have found many results that conflict with this assumption. For example, as mentioned earlier, for a hybrid category learning task in which a simple rule is unable to produce maximum accuracy, people tend to use a simple rule to generate responses (Ashby and Crossley, 2010). Paul and Ashby (2013) showed that the behaviour of COVIS appears to be very similar for both II and hybrid categories, whereas in the simulation study reported here hybrid category learning is not controlled by the procedural learning system. In addition, at the algorithmic level, COVIS seems to predict that trial-by-trial system switching is impossible. (Though neuropsychologically, COVIS assumes that trial-by-trial system switching is a common occurrence, it is not reflected in the computational implementation.) This prediction is also problematic. As mentioned in Chapter 2, several studies have shown that control of systems is switchable on a trial-by-trial basis (e.g., Crossley et al., 2017; Erickson, 2008; Helie, 2017; Paul & Ashby, 2013). Nevertheless, while the existing dual systems mechanism is robust, there must also be a control mechanism for interaction between multiple representations. It appears that this control mechanism must be more complex than system weights combining trust with confidence.

As a CCN model, COVIS has many important advantages. First, rather than making predictions about purely behavioural dependent measures (i.e., accuracy and response time), COVIS is able to make predictions about other types of data, including fMRI data, EEG data, and single-cell recordings. Second, the neurobiologically-constrained approach is easier to be falsified, and therefore may facilitate theoretical work. For example, COVIS holds that the striatum is critical to category learning. The neuroscience result is now well-established that any theory of category learning that attends to neuroscience must address the importance of the

striatum. Since the neuroanatomy of the striatum is well understood, along with its major inputs and outputs, this means that any CCN model of category learning must converge on a similar architecture. More details might be added, and a somewhat different computational role might be assigned to certain components, but it seems unlikely that this basic architecture will disappear. Third, the model provides a clear illustration of the dual system framework.

However, it is important to note that the CCN modelling requires a good understanding of the cognitive processes (mechanisms and algorithms). The CCN approach should not replace mechanistic modelling. Instead, the CCN approach and mechanistic modelling are complementary. For building up a neurobiological model, a more complete understanding of the behaviour is required. To this end, mechanistic modelling results are critical to the process of building up the knowledge base needed to develop a CCN model. However, so far, the fact remains that no complete mechanistic model based on the dual systems framework exists. Although there is plenty of neuroscience data to support a dual systems framework, we do not know much about the neural mechanisms mediating interactions between multiple representations. There have been a few behavioural studies supporting the existence of a control mechanism. But, the CCN approach may be premature to model this data. Rethinking of the algorithms is more important and straightforward. Progress at the algorithmic level may also facilitate the extension of COVIS. It thus seems appropriate to revisit the mechanistic, modular network model, ATRIUM.

5.4.2 Exemplar-Based Representation and Attention

COVIS postulates that only the rule-based system is mediated by selective attention, whereas the exemplar-based system is independent of attention. This is where the CCN model and current attention learning models diverge. The most powerful evidence supporting the attentional processes independent hypothesis comes from a series of early studies of dual task interference effects on different category learning tasks (e.g., Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). In Waldron and Ashby (2001), participants were given a category learning task using geometric patterns which could vary on four binary dimensions (e.g., shape: circle or triangle). In the II task, three out of four of these dimensions were relevant

to the classification rule, and in the RB task, only one dimension was relevant. Participants in the dual task conditions performed a numerical Stroop task concurrently with the category learning task. Participants performing the RB in the dual task condition took longer (more trials) to reach criterion than those performing the RB task alone, but dual task interference had a negligible effect on II learning. Zeithamova and Maddox (2006) replicated this effect using continuous category structures.

However, there have been much recent evidence against the early evidence for the attentional processes independent hypothesis. Nosofsky and Kruschke (2002) demonstrated that the ALCOVE model could naturally predict the behavioural pattern observed by Waldron and Ashby by suggesting that the dual task interference impaired participants' ability to attend selectively to relevant stimulus dimensions. Zeithamova and Maddox's (2006) replication of the selective effect of dual task interference on RB tasks has also been challenged. Newell et al. (2010) failed to replicate the selective effect in three experiments using a variety of dual task conditions. They concluded that Zeithamova and Maddox's original interpretation had been confounded by the inclusion of 'non-learners' in the analysis (i.e., those participants who had neither learned the category task, nor performed the secondary task adequately). Once these participants were removed, all evidence for a selective effect of dual task interference on RB and II tasks disappeared. Miles et al. (2014) further reported that when the secondary task continuously taxed attentional processes, neither RB nor II categories could be learned appropriately.

One of the most popular II tasks is known as the weather-prediction task (WPT) (Knowlton et al., 1994). In the original WPT, one, two, or three of four possible tarot cards are shown to the participant, whose task is to indicate whether the presented constellation signals rain or sun. Each card is labelled with a unique, and highly discriminable, geometric pattern. Fourteen of the 16 possible card combinations are used (the zero and four-card combinations are excluded) and the optimal strategy requires using all available cues. Knowlton et al. (1996) reported that patients with Parkinson disease showed comparable performance to aged-matched controls, whereas amnesic patients performed more poorly than controls. Ashby et al. (1999) used Knowlton et al.'s work as important evidence for the attentional processes independent hypothesis. However, a preliminary behavioural study reported in Appendix A, using the WPT

with a range of auditory-vocal secondary tasks held to tap various attentional processes, found that all secondary tasks significantly interfered with WPT performance, again, suggesting that absence of attentional processes prevents II category learning.

In fact, the exemplar-based representation need not necessarily be independent of attentional processes. Many researchers have recently argued that attentional processes may participate in mediating the competition between multiple representations (e.g., Ashby & Crossley, 2010; Erickson, 2008; Miles et al., 2014; Paul & Ashby, 2013). For example, Ashby and Crossley (2010) proposed that, though the frontal cortex does not directly affect decision making in II category learning, it connects via the hyperdirect pathway to the subthalamic nucleus, which in turn can affect activity in the basal ganglia and the cortex. The frontal cortex can increase subthalamic activity, which inhibits communication between the basal ganglia and cortex. Since responses from the exemplar-based representation system are generated in the basal ganglia, decreased communication between the basal ganglia and the premotor/motor cortex impedes the execution of responses from the exemplar-based representation system. Therefore, the frontal cortex's ability to inhibit communication between the basal ganglia and the cortex could be the mechanism that mediates competition between multiple representations (see Chapter 7 for more detail).

5.4.3 Modular Architecture and Heterogeneous Category

Representations

The modular architecture of the original COVIS model was designed to account for behaviour on homogeneous category learning tasks, but not for that on heterogeneous category learning tasks. In a homogeneous category learning task, as the category representation is integrated and exclusive, participants' performance can be explained by the competition between multiple representations. In contrast, in a heterogeneous category learning, task partitioning and switching between strategies (representations) are essential. Thus, the system competition mechanism of COVIS is insufficient. A higher-level cognitive process seems to be necessary in the modular architecture of heterogeneous category representations (e.g., Ashby & Crossley, 2010; Erickson, 2008).

Chapter 6.

Case Study III: The Modular Architecture of ATRIUM and the Mechanism of Representational Attention

6.1 Introduction

6.1.1 The Modular Architecture of Category Learning

As mentioned in previous chapters, multiple investigations of context-dependent partitioning of knowledge about category boundaries have revealed that the context in which a stimulus is presented can modulate the attention paid to it (e.g., Aha & Goldstone, 1992; Lewandowsky et al., 2006; Sewell & Lewandowsky, 2011; Yang & Lewandowsky, 2003; 2004). As shown by Erickson and Kruschke (1998), the learned influence of context-dependent representation is relatively long term, extending even into transfer which involves novel stimuli (e.g., George & Kruschke, 2012).

Studies such as these suggest that people are able to form category representations that are composed of multiple subsets of stimuli. This modular architecture of category representation/sub-representation structure shares some properties with Norman and Shallice's (1986) schema theory of action control (see Chapter 7 for detail). In the Norman-Shallice model, a schema refers to a representation of an action or a thought. In addition, a schema can be divided into several subschemas. If the whole modular architecture of category representation is regarded as a schema, then associations of subset can be thought to be subschemas. As the

relationships between these subsets are heterogeneous, the formation of category representations requires a similar selection mechanism to that of the organisation of hierarchical schema/subschema structure, that can facilitate switching between different subsets.

Unfortunately, though the schema theory does provide a good account for the hierarchical organisation of heterogeneous representations, there has been no attempt, to my knowledge, to explore where the relationships of the hierarchy of schemas come from. How can we learn the hierarchical structure of the category representation? I suggest that attention should play an important role in this process. Thus, this Chapter presents one of the most prominent computational accounts in the categorisation literature, ATRIUM (i.e., Attention to Rules and Instances unified model), in an effort to illustrate the linkage between attention and the organisation of internal representations in category learning.

6.1.2 The Mixture-of-Experts Network

ATRIUM uses a mixture-of-experts approach. This refers to a machine learning technique in which multiple local solutions (expert networks) are used to partition the whole problem space into independent sub-task regions (Jacobs et al., 1991a; 1991b). The expert networks compete to learn. The competition is coordinated by a gating network. The gating network learns to allocate the expert network whose solution is the most appropriate to each sub-task.

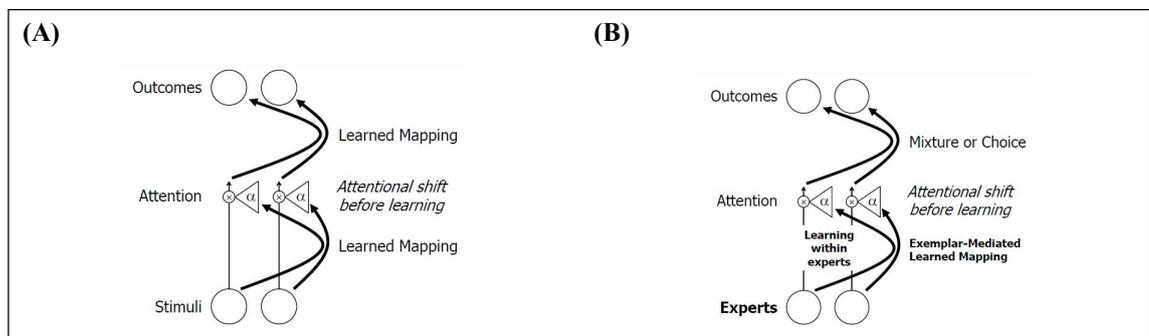


Fig 6-1. Schematic representation of the structure of (A) the standard exemplar-based network, and (B) the mixture-of-experts architecture. In the modular architecture, attention is allocated to expert modules, each of which constitutes a complete learning system from stimuli to outcomes. The mapping from stimuli to attention is exemplar-based, so that different experts can dominate in different contexts. The internal details of the experts are the same as the standard exemplar-based network.

As a single system network model (ALCOVE) is unable to account for multiple representations. To deal with multiple representations in category learning, one possibility is to instantiate a modular architecture which partitions the task into sub-tasks. Within such an architecture, on one hand, each module can learn independently and compete to produce output. On the other hand, selection among modules should be exemplar-based. The idea of an exemplar-based modular architecture comes from research on heterogeneous categorisation. For example, Lewandowsky and his colleagues (Little & Lewandowsky, 2009; Sewell & Lewandowsky, 2012; Yang & Lewandowsky, 2003; 2004) have confirmed that people tend to make categorisation decisions based on exemplar-similarity. This approach can also account for the phenomenon of rule-based extrapolation (Denton et al., 2008; Erickson & Kruschke, 1998; Kruschke & Erickson, 1994).

Fig 6-1B shows the basic schematic representation of a mixture-of-experts architecture that can fit both principles. A mixture-of-experts network can have several expert modules, where each module learns its own mapping from stimuli to outcomes using its own form of representation. The output and learning of these sub-representational modules are determined by a gating network, which learns to activate a sub-representational module in response to a particular input stimulus.

In the literature, three versions of ATRIUM have been formally proposed. The first version is the Kruschke and Erickson (1994) version. In that version, the focus is on addressing learning of rule-plus-exception tasks. The use of the gating network reflects the mechanism of coordinating the competition between a local rule module and a local exemplar module. The second version—the Erickson and Kruschke (1998) version—was used to fit rule-based extrapolation effects. In order to fit human data, Erickson and Kruschke (1998) introduced a dimensional attention mechanism into the exemplar module, which was not considered in their 1994 version. But, it causes two issues.

First, Kruschke and Erickson (1994) did not apply the dimensional attention mechanism in the exemplar-based representation module as they suggested that the invoked local rule module could reflect the function of dimensional attention, but the 1998 version applied dimensional attention in the exemplar-based representation module which invoked the whole implementation of ALCOVE. Thus, the 1998 version indeed has invoked two types of dimensional attention.

However, it is problematic because no attempt is made to explain the overlapped attentional function in one model.

Second, as some have proposed, ALCOVE as a single module model can capture learning of either rule-based or exemplar-based tasks (e.g., Nosofsky et al., 1994). Therefore, in the second version, the exemplar module can eventually dominate the task via learning. However, there is recent evidence that rule-based and exemplar-based tasks are represented differently even at the automatic level. Specifically, it is proposed that in the premotor cortex, explicit and implicit representations are associated with different neurons (see Chapter 2).

Lewandowsky et al. (2006) introduced the third version of ATRIUM which is used to model knowledge partitioning in categorisation. The architecture of the third version is the same as the previous ones, except that the authors added one more rule module (see Erickson & Kruschke, 2001, page 2), and changed the gate node activation (or probability to select a local module).

This chapter focusses on the 1998 version of ATRIUM, because Erickson and Kruschke (1998) provide relatively more implementation details, and because this version seems to be more influential in the literature. The next section describes a reimplementing of the mixture-of-experts model. The model description and implementation follows Erickson and Kruschke's (1998) paper.

6.2 Reimplementing ATRIUM

6.2.1 Model Description

By combining the mixture-of-experts approach with an error reduction mechanism, Erickson and Kruschke (1998) proposed the ATRIUM model. In ATRIUM, a gating network learns to allocate responsibility to representational modules, just like attention allocation. The gating network is also referred to as *representational attention* (see Fig 6-2). Learning in ATRIUM is driven by error reduction. Attention is allocated to representational modules that accommodate the current training case, and away from modules that cause mistakes. The gating

network is an exemplar-based mapping. In other words, in this account, instead of error-driven dimensional relevance shifts, selection of the appropriate subset of task knowledge is based on exemplar similarity.

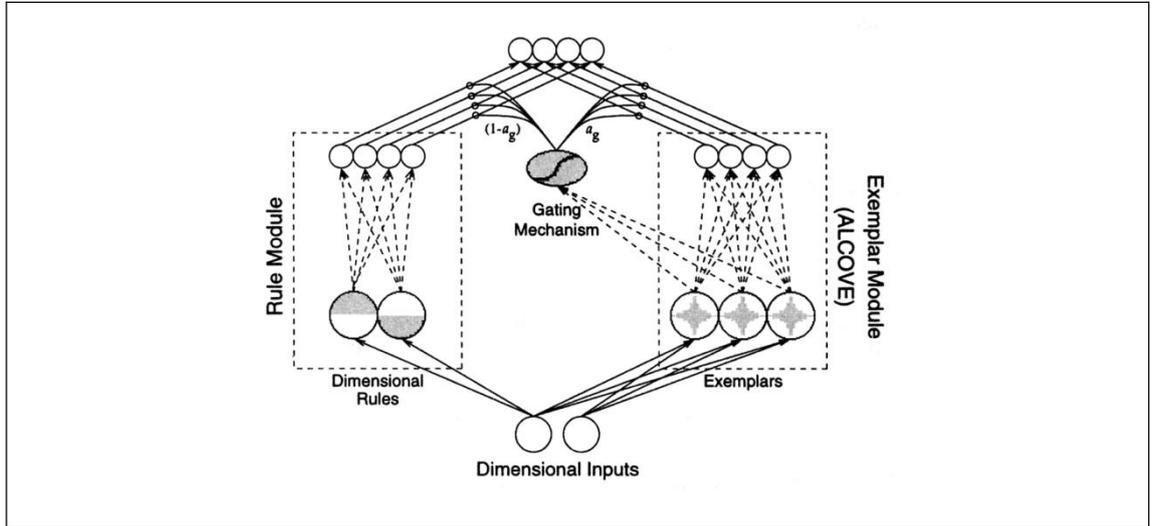


Fig 6-2. Schematic representation of the architecture of the original version of ATRIUM (From ‘Rules and Exemplars in Category Learning’ by M. A. Erickson and J. K. Kruschke (1998). Journal of Experiment Psychology. 127(2). pp 117. Copyright 1998 by American Psychological Association, Inc).

6.2.1.1 Rule-Based Module

Rule modules represent decision boundaries that bias the category space into different response regions. Instead of rule selection processes, learning in the rule-based module is associative learning. Each rule module includes a pair of rule nodes whose activation is determined by the positioning of the stimulus input relative to the boundary represented of that module. The extent to which rule nodes (denoted by + and – subscripts, respectively) are activated is given by a sigmoid function:

$$a_{r+} = \frac{1}{(1 + \exp(-y_r(a_i^{in} + \beta_r)))} \quad (6.1)$$

$$a_{r-} = 1 - a_{r+} \quad (6.2)$$

where y_r , the gain of the sigmoid, is proportional to the standard deviation of the perceptual or criterial noise associated with the rule, thus determining the rule’s ‘sharpness’, and β_r determines the position of the rule boundary. During training, response regions on either side of the rule boundary are associated with category output nodes by gradient descent on error. The

rates at which the rule module learns the rule-node-to-category-node association weights are determined by rule learning rate λ_r . (See Erickson and Kruschke, 1998 for details.)

6.2.1.2 Exemplar-Based Representation Module

The exemplar module is a standard implementation of ALCOVE (see Chapter 4) except that instead of being used to directly compute responses, activation of the category nodes are fed forward to the gating network as candidate responses.

6.2.1.3 Gating Network

As mentioned above, the function of the gating network is to establish the mechanism for exemplar-based representational attention shifts (Jacobs, 1997; Jacobs et al., 1991). The exemplars represented in the gating module are the same as in exemplar module. The association between each exemplar node and gate node is a function of the accuracy achieved by the module. The output of a gate node is computed in two steps. In the first step, each net activation, a_{gm} , is computed by summing the weighted activation of all the exemplar nodes. Thus,

$$a_{gm} = \sum_{ej} w_{gm,ej} a_{ej} \quad (6.3)$$

where $w_{gm,ej}$ is the association weight from exemplar node j to gate node m . Erickson and Kruschke (1998) argued that the activation of gate nodes must be squashed in the range of $[0,1]$ to represent the probability of using each module. Thus, the probability of choosing the exemplar module is given by:

$$p_{ex} = \frac{1}{(1 + \exp(-y_g a_{gex} + \beta_g))} \quad (6.4)$$

and the probability of choosing the rule module is given by $p_r = 1 - p_{ex}$. In tasks which involve more than just two types of representations, equation 6.4 can be transformed as:

$$p_m = \frac{\exp(y_g a_{gm} - \beta_{gm})}{\sum_m \exp(y_g a_{gm} - \beta_{gm})} \quad (6.5)$$

where, β_g is the gate bias, and $y_g \geq 0$ is the gate gain. If y_g is high, differences among the gate node activations are enhanced in the gate node probabilities. If y_g is low, differences among the gate node activations are attenuated in the gate node probabilities.

Activation levels of the category nodes in each module are fed forward to the gating network as candidate responses. The final output of ATRIUM, the probability of choosing category K , $P(K)$, is given by:

$$P(K) = p_m \frac{\exp(\phi \alpha_k^m)}{\sum_k \exp(\phi \alpha_k^m)} \quad (6.6)$$

where $\phi \geq 0$ is a scaling constant, which may be thought of as representing the level of decisiveness in the system. If ϕ is high, differences in activation are accentuated. If ϕ is low, differences in activation are diminished in the final output.

6.2.2 Modelling Erickson and Kruschke (1998) Experiment 1

6.2.2.1 Introduction

ATRIUM, as an account of exemplar-based attention to representations, has exhibited several advantages in categorisation. Firstly, adding explicit rule learning makes the model learn faster in rule-based category learning compared to models based on only a dimensional relevance shifting mechanism (Erickson & Kruschke, 1998). Secondly, incorporation of multiple representations provides much more flexibility in accounting for variation of heterogeneity of category structures. Indeed, incorporation of an exemplar-mediated gating network uses the mechanism of selection of stimulus-dependent representations to replace the selection of stimulus-dependent responses. It is consistent with the assumption that people should attend to the appropriate strategy before an appropriate response is produced at any moment. In a sense, ATRIUM provides a unified account of the formation of task-dependent representations in categorisation. In other words, the learning process of this model can be thought of as an interpretation of the process of the formation of a schema in categorisation.

ATRIUM was designed not only for multiple representations in category learning, but also for the formation of heterogeneous representations. According to ATRIUM, category learning is

mediated by at least two types of representations: rule-based and exemplar-based representations. More importantly, ATRIUM assumes that selection of a response is mediated by an exemplar-based representational attention mechanism rather than the dimensional attention mechanism. Through the gating network, ATRIUM can appropriately decide which representation to select corresponding to the situation (exemplar similarity) rather than which dimension to focus on.

6.2.2.2 Method

For fitting the human data of the rule-plus-exception category learning, the implementation of ATRIUM follows the model assumptions described earlier in this chapter (see also Erickson and Kruschke, 1998, for more detail). The programming here used MATLAB. The model was run for 100 simulations. In each simulation, the model was first trained over the course of 29 blocks of 14 trials. In each block, stimuli were presented in a random ordered sequence. After training, the learned attention strengths and association weights were then applied to fit to the transfer stimuli in a random order. This is analogous to the procedure in the empirical manipulation of Erickson and Kruschke (1998) Experiment 1. The replication here, as was done by Erickson and Kruschke (1998), has shown the power of the model in the prediction of human learning and generalisation. I used the same parameter settings as estimated by them.

6.2.2.3 Results

Fig 6-3 top panel shows the fit of ATRIUM to the training data. As can be seen in the figure, not surprisingly, ATRIUM is able to provide a very good fit to the data. According to Erickson and Kruschke (1998), accuracy in learning rule stimuli improved from 28% (ATRIUM: 27.5%) to 87% (ATRIUM: 86.6%), while the exception responses to those stimuli decreased from 25% (ATRIUM: 24.4%) to 5% (ATRIUM: 4.1%). In contrast, the accuracy for learning exception stimuli improved from 27% (ATRIUM: 25.4%) to 86% (ATRIUM: 86.1%), while rule responses to those stimuli decreased from 29% (ATRIUM: 30.1%) to 12% (ATRIUM: 10.2%). (The pattern of learning curves can also be found in Erickson and Kruschke, 1998.)

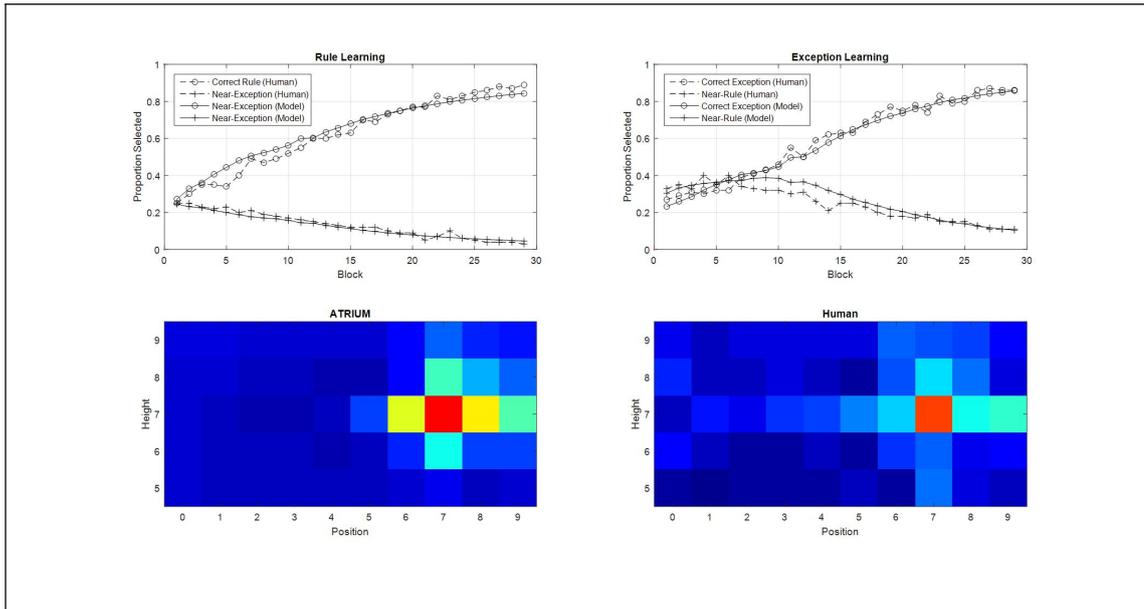


Fig 6-3. Top: The ATRIUM’s prediction of proportion of rule and exception responses as a function of learning blocks. Bottom: The proportion of exception responses in the test phase observed human participants’ data in Erickson and Kruschke (1998) Experiment 1 and ATRIUM’S prediction. Cool colours represent low choice probability whereas warm colours represent high choice probability.

As mentioned in Chapter 3, though the ALCOVE model also did a good job in fitting the training data, it failed to predict the proportion of exception responses observed in Erickson and Kruschke’s (1998) experiment (see Fig 3-4). In contrast, ATRIUM was able to account for rule-based extrapolation. The bottom panel of Fig 6-3 shows the proportion of exception responses observed in Erickson and Kruschke’s (1998) experiment and predicted by ATRIUM. As can be seen in the Figure, ATRIUM qualitatively fit to the observed transfer data. In particular, ATRIUM’s prediction of the proportion of exception responses on the stimulus T_E (see Fig 3-3A) is 15.5%, which is close to the observed data (11%). This performance is better than ALCOVE which predicted that the proportion of exception responses on the stimulus T_E is 42.4%. In addition, Erickson and Kruschke (1998) reported that, in the transfer phase, participants were more likely to give exception responses to stimuli which matched the exception on the primary dimension. ALCOVE failed to predict this effect. But, as can be seen in the bottom panel Fig 6-3, ATRIUM has successfully predicted that the proportion of exception responding is greater on those that matched the exception on the primary dimension than those that matched the exception on the secondary dimension.

As the reimplementation of ATRIUM can quantitatively fit the data, and as the fitting results are almost the same as shown in Erickson and Kruschke (1998), it can be concluded that

the replication is successful. However, the target in this section is not merely to reproduce the quantitative fit of the model to the data, but also to exemplify the idea of exemplar-based attention to representations. Fig 6-4 demonstrate how each part of the mixture-of-experts network learns. According to Erickson and Kruschke (1998), the gradient error in ATRIUM can be converted to the accuracy. The accuracy of each module m is defined as

$$A_m = \exp\left(-\frac{1}{2} \sum_k |t_k^m - a_k^m|^2\right) \quad (6.7)$$

The mean accuracy of the model is then expressed

$$\bar{A} = p_m A_m \quad (6.8)$$

so that the total error can be expressed as $E = -\log(\bar{A})$. As can be seen in Fig 6-4, since the performance of the rule module is not stable, the gating network gradually learns to bias processing to the exemplar-based representation module.

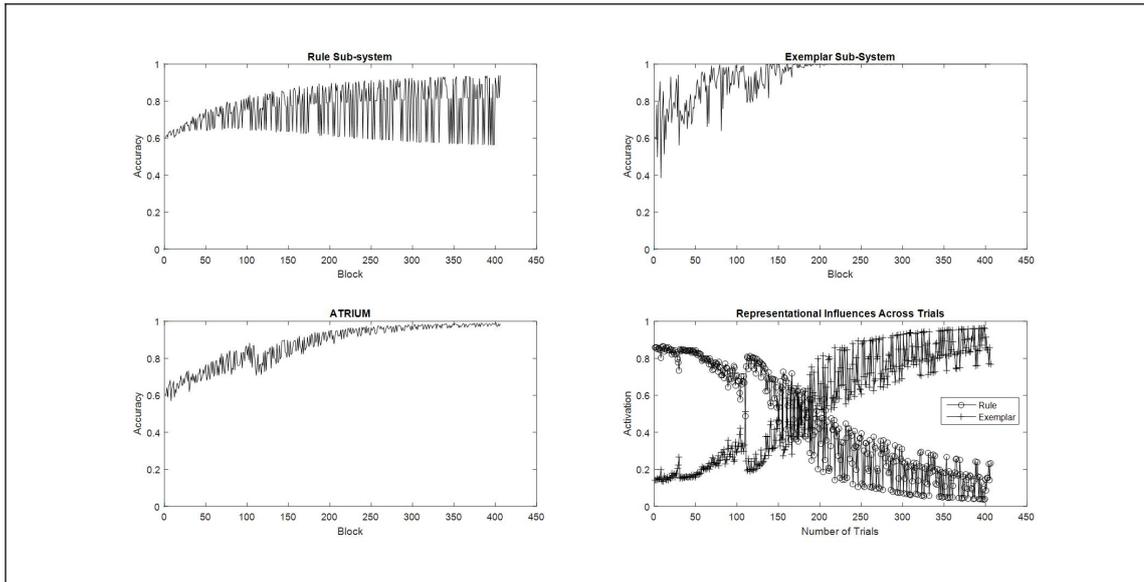


Fig 6-4. An illustrative example of how ATRIUM, its sub-systems and gating network learn via gradient descent on error across 29 training blocks.

Fig 6-5A shows the profile of learned exemplar-based representational attention strengths (connection weights between exemplar nodes and the gate node). The original version of ATRIUM just incorporated one gate node, thus the weights in this profile indicate the attention strengths to exemplar-based representations. As can be seen in the Figure, great attention strengths are assigned to the training exception (e.g., 2.0620 for stimulus [2,2] and 2.0910 for

stimulus [7,7]), whereas other exemplars achieved relatively low attention strengths. This profile, thus, reveals that the success of ATRIUM to reproduce the rule-based extrapolation is attributed to the constraint of attention focused on specific exemplars. Moreover, Fig 6-5B shows that the exemplar-based representation module of ATRIUM learned to pay greater attentional strength on the primary dimension than the secondary dimension, whereas ALCOVE did not. This also facilitates the gating network to learn to form the rule and exception representations.

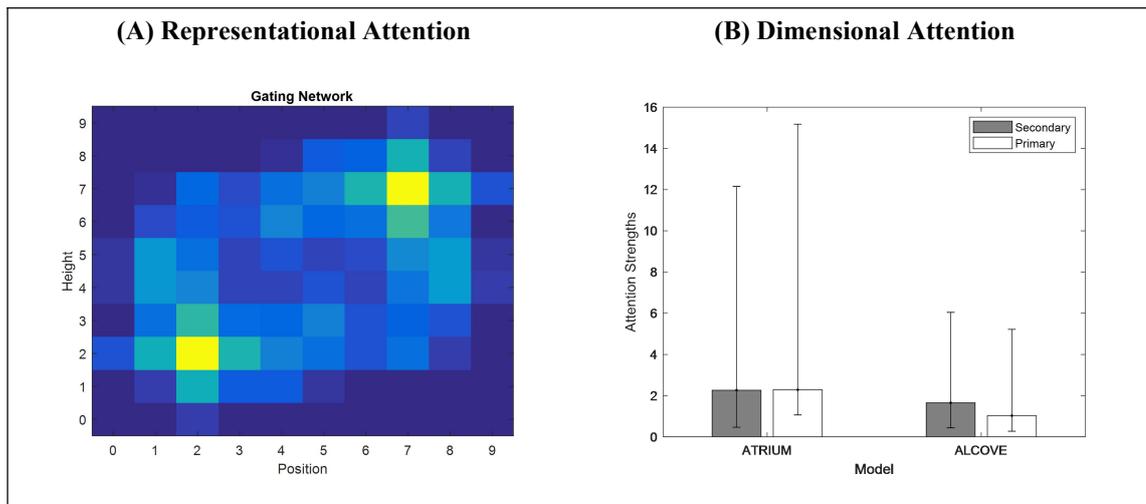


Fig 6-5. Panel (A): The profile of learned exemplar-based attention produced by the gating network for achieving the rule-plus-exception task. Panel (B): The learned dimensional attention (error bars represent maximum and minimum values across 100 stimulations).

6.3 Simulation Study: Hybrid Category Learning

Although ATRIUM is inspired by the multiple representations assumption and shares some similarity with COVIS, it has never been applied to the large-scale, probabilistic category structures. ATRIUM has been argued to be advantageous in accounting for deterministic and heterogeneous category structures (e.g., Nosofsky & Little, 2010; Sewell & Lewandowsky, 2011; Yang & Lewandowsky, 2006), yet no attempt appears to have been made to examine the capacity of the model to simulate II and Hybrid category learning. However, establishing these results seems to be a necessary first step in determining whether the principles of ATRIUM can be considered as a good candidate account of interaction between distinct representational systems. This Section demonstrates how the mixture-of-experts approach of ATRIUM accounts for II and hybrid categories (Paul & Ashby, 2013).

6.3.1 Category Structures

In this simulation, the set of hybrid categories and the set of II categories shown in Fig 3-2 were simultaneously used to train the model. For the II categories, a total of 300 stimuli were drawn from each of two bivariate normal distributions. The category means and variances are identical to the II categories used in Chapter 5. The suboptimal one dimensional rule can only reach 78.5% accuracy. For the hybrid categories, a total of 300 stimuli were randomly sampled from each of two categories separated by a hybrid boundary. The hybrid category structure is, too, identical to that of Chapter 5. The maximum performance by the suboptimal one-dimensional rule on these categories is about 90%. Note that, in this simulation, the decision boundary in the rule module is set to achieve the suboptimal performance for both II and hybrid category learning tasks.

6.3.2 Method

The implementation of ATRIUM in this simulation is identical to that of Erickson and Kruschke (1998). As mentioned above, the mixture-of-experts architecture consists of a rule-based module, an exemplar module and a gating network. The rule in the rule module is represented as the middle point of a single dimension. For simplicity, the hidden layer of the exemplar module includes 2500 exemplar nodes arranged in a 50 x 50 grid (because this simulation uses stimuli that vary in two dimensions).

As ATRIUM is sensitive to stimulus orderings, it is necessary to have a fixed random sample on which to run this simulation. For this reason a random set of 15 random stimulus orderings was generated. Each step of this simulation was tested on this training set, and the model performance was averaged across them. This thus provides stable estimates of the model behaviour. In addition, it is important to select values for the parameters defining the parameter search of the simulation. Six parameters are needed: the four learning rate parameters, the

exemplar specificity and the gate bias parameter. Table 6-1 summarises the function and search range for each selected parameter. Every other parameter was fixed to a specific value ($\phi = 4.0$; $y_r = 0.9$; $y_g = 1.0$).

This simulation was run on 10,000 steps. 10,000 points were randomly sampled from the parameter space by the Monte Carlo method. Each step consisted of 15 blocks of II and hybrid category learning including 600 trials with re-initialisation of the model occurring between blocks.

Table 6-1.
Function and search range for ATRIUM's six parameters in the simulation study

| Parameters | Function | Range |
|-------------|-------------------------------------|------------|
| c | Specificity of exemplar nodes | [0.2, 2] |
| η_{ex} | Learning rate of exemplar module | [0.01, 15] |
| η_a | Dimensional attention learning rate | [0.01, 15] |
| η_g | Learning rate of gating network | [0.01, 15] |
| η_r | Learning rate of rule module | [0.01, 2] |
| β_g | Bias for the exemplar module | [0.5, 3] |

The output of the model is the choice probability of correct response (accuracy). The results of 600 trials are summarised as an array of 6 numbers. Each number indicates the mean accuracy averaged across each 100 trials. The mean accuracy of the last 100 trials is used as the measure of the model performance, because in a typical empirical study the period of the first 500 trials is often considered as the training phase, while the last 100 trials are considered as the transfer trials. This definition of model performance easily allows for quantification of how well the model performs on the task after training.

6.3.3 Results and Discussion

Fig 6-6 shows the distribution of model performance from the 10,000 Monte Carlo simulations. As the parameter values change, ATRIUM produces many possibilities. The range of proportion correct for the II task is predicted as between 68.89% and 98.06%, whereas for the hybrid task it is between 71.80% and 95.26%.

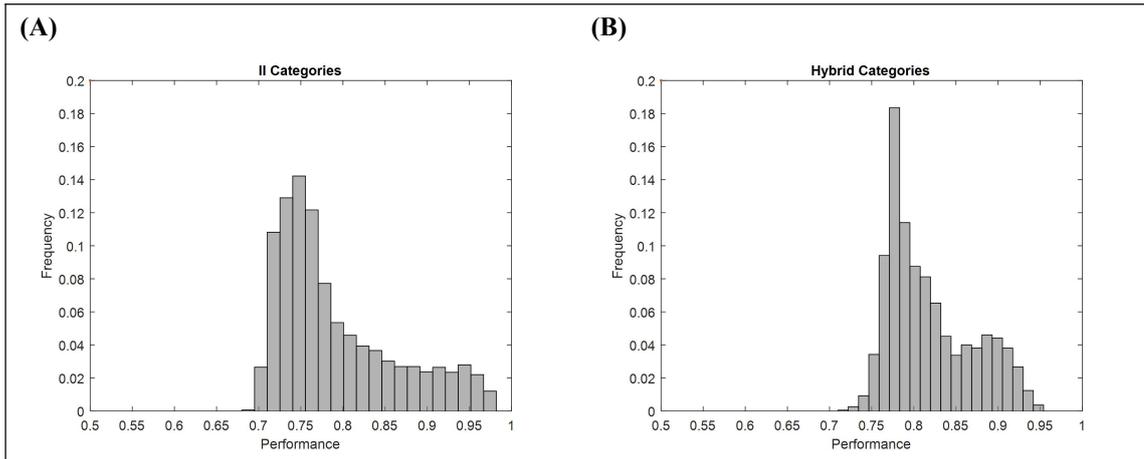


Fig 6-6. The distribution of observed ATRIUM's performances (proportion correct) on learning II categories (A) and hybrid categories (B). The data shown is a random sample of 10,000 points from the parameter space.

For the II categories, although there is no test performance below 65%, only 12.20% of the volume of the 10,000 points learned the categories at or above 90% accuracy. 65.39% of the volume falls into the range between 70% and 79% accuracy, and 22.40% falls into the range between 80% and 89% accuracy. For the hybrid categories, only 9.95% of the volume learned the categories at or above 90% accuracy. 42.90% of the volume falls into the range between 80% and 89% accuracy, and 47.15% falls into the range between 70% and 79% accuracy.

It seems that ATRIUM can do a good job at learning II and hybrid categories. However, a more important question is how does ATRIUM learn these category structures, or what does the model learn? Fig 6-7 shows the learned gating network weights at the end of training. As can be seen in the figure, obviously, ATRIUM provides a strong suggestion of trial-by-trial switching between different representation modules. However, according to Ashby and Crossley (2010), learning in II and hybrid categories should be controlled by different representation modules. II category learning should be controlled by the exemplar-based module, whereas for hybrid category learning, the rule module should be in control. However, it seems that ATRIUM is unable to balance the strategy use in these two tasks. In particular, although, ATRIUM can predict the II category learning controlled by exemplar-based module, it cannot ignore the influence from exemplar-based representations on the left corner in the hybrid category structure (see Fig 6-7A and Fig 6-7B). In contrast, when the model ignores the influence of the exemplar-based representation in hybrid category learning, it cannot produce II category learning controlled by exemplar-based representation (see Fig 6-7C, D, E and F).

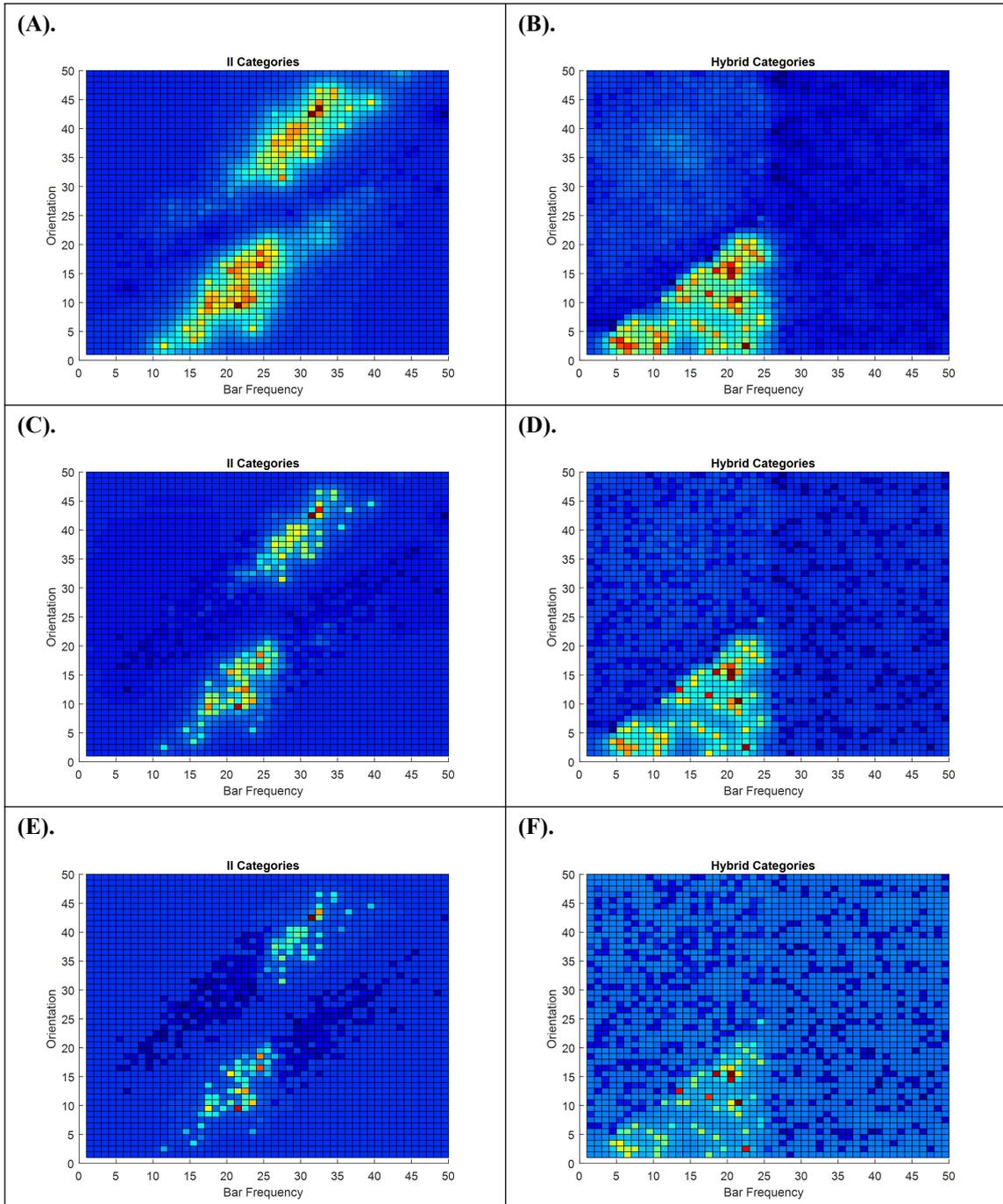


Fig 6-7. Top: Weights in the gating network trained on II categories (A) and hybrid categories (B) averaged across the performances above 90% accuracy. Middle: Weights in the gating network trained on II categories (C) and hybrid categories (D) averaged across the performances ranged between 80% and 89% accuracy. Bottom: Weights in the gating network trained on II categories (E) and hybrid categories (F) averaged across the performances ranged between 70% and 79% accuracy. Blue (cool colours) represent small weights whereas red (warm colours) represent large weights.

It is not surprising that ATRIUM seems unable to balance the strategy use in the tasks. Because, as mentioned earlier, ATRIUM was originally designed to account for heterogeneous category representation, the learning algorithm was intentionally designed to facilitate the optimisation based on trial-by-trial switching between different representations. For hybrid

category learning, the rule module cannot provide the optimal solution for those left corner stimuli, but the exemplar-based representation module can. Of course, the gating network learns to involve the representational influence of the exemplar-based module into the hybrid category learning.

This simulation suggests that ATRIUM predicts an easy trial-by-trial system switching in hybrid category learning. However, although Erickson (2008) (and also Crossley et al., 2017; Helie, 2017) suggested that trial-by-trial switching between modules is possible on a heterogeneous, hybrid category structure, Ashby and his colleagues suggested that, for a homogeneous, hybrid category structure, human participants tend to use a rule-based strategy (e.g., Ashby & Crossley, 2010; Paul & Ashby, 2013). This disagreement between ATRIUM and COVIS leads to a contradictory interpretation of the control of multiple internal representations in human category learning. The putative trial-by-trial switching mechanism in the gating network has a certain degree of rationality, given that there is a more advanced control mechanism to constrain it. Given that the trial-by-trial switching mechanism reflects a bottom-up process, there should be a higher-level, top-down control that determines the interaction between different representations.

6.4 Representational Attention and Task Switching

6.4.1 Motivation

ATRIUM was originally designed for rule-based extrapolation effects in rule-plus-exception category learning (see also Denton et al., 2008). In a preliminary case study we found that the mixture-of-experts architecture is also suited to account for the Aha and Goldstone (1992) data set (see Appendix C). Previous computational modelling studies also revealed that ATRIUM did a good job at predicting human generalisation performance in the task partitioning paradigm in which stimuli are presented with a binary contextual cue (e.g., background colour) to signal the presence of a local regularity in heterogeneous category representations (e.g., Sewell & Lewandowsky, 2011; 2012; Yang & Lewandowsky, 2004).

However, the stimuli used in the earlier experimental paradigms in heterogeneous category learning are sampled from a small number of exemplars that are deterministically assigned to the alternative categories, but not large-size, probabilistic categories.

Erickson (2008) first combined the traditional heterogeneous category learning paradigm (e.g., multiple phases of training and extra contextual cues) with large-size, probabilistic category structures. As mentioned in Chapter 2, the most important contribution of Erickson's (2008) experiment was the discovery of task switching effects in heterogeneous category learning (see Section 2.4.3 and Fig 2-9). Erickson (2008) also argued that the modular architecture of ATRIUM could be easily modified to approach the task switching paradigm. But, unlike traditional paradigms, the focus of task switching paradigms is not on choice probability of specific exemplars to certain categories but switch costs (e.g., reaction times) when selecting between internal representation modules. Therefore, it is longer necessary to retain some properties of the model.

There are three issues in modelling Erickson's (2008) data set. The first issue is that the attention learning mechanism of the original version of ATRIUM may be not required. In traditional experimental paradigms, the attentional focus on each dimension was unknown for participants in the beginning of training, and thus attention learning is needed to model the data. But, in the Erickson (2008) experiment, the dimensional attentional focus was told to subjects before training, meaning that attention strengths on stimulus dimensions could be set as constants rather than adjusting by error-driven learning.

Second, the form of 1D rule-based representation in the original model may be too complex for task switching. In the original version of ATRIUM, the rule module indeed learns to associate 1D rule-bounded regions of psychological space with category selection (see Fig 6-8B). This account is plausible for deterministic categorisation tasks, in particular for generalisation performance. However, in Erickson's experiment, there are two subtasks, and thus, different tasks may involve different alternative decision bounds to compete with the exemplar module. In other words, according to the idea of the original ATRIUM model, there might be, at least, three internal modules (i.e., two rule modules and one exemplar module) competing to dominate the final output in the switching between two tasks. For simplicity, we assume that there are only two modules in task switching (i.e., a 1D module and a 2D module), and use an exemplar

network to replace the decision bound network, but the dimensional attention strengths are set to be different in the different modules. Thus, the attention strengths in the 1D module are set as 0 on rectangle height and 1 on segment line position, whereas in the 2D module, the attention strengths are set as 1 for both stimulus dimensions. Some may argue that this 1D representation module is not suitable for accounting for generalisation, but it is enough for task switching.

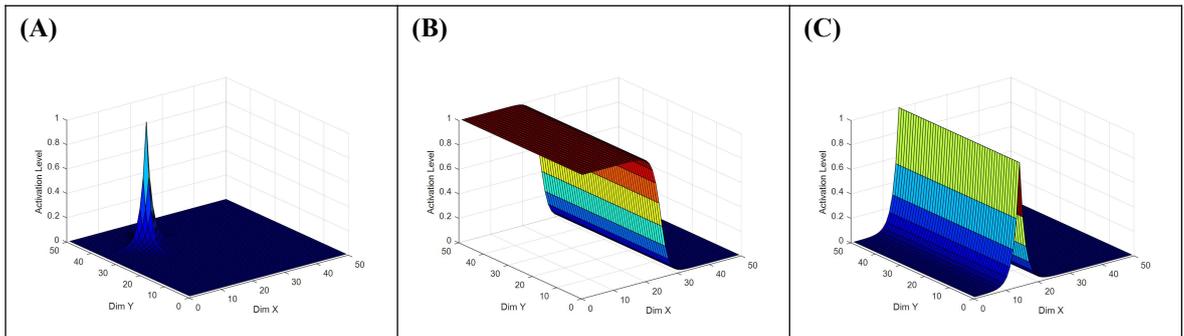


Fig 6-8. Schematic illustration of activation profiles in exemplar module (A) and rule module (B) on an example single trial according to the original ATRIUM model, and activation profile in 1D module of the modified model (C).

The third noteworthy issue is about the role of contextual cues in category learning. In Erickson's experiment, in addition to two stimulus dimensions, there is an extra binary contextual cue (i.e., background colour). How can we interpret the role of contextual cues? More recent evidence reveals that the contextual cue is more likely to play a critical role in driving higher-level attentional control (e.g., Ashby & Crossley, 2010; Crossley et al., 2014; 2017; George & Kruschke, 2012; Sewell & Lewandowsky, 2011; 2012). This is in line with the tradition of task switching theories, because in task switching paradigms, a number of cues are always included that signal which task representation should be applied. Therefore, the position in this modified model is that the binary contextual cue should receive greater strength than the stimulus dimensions.

Although compared with other models (e.g., SUSTAIN and COVIS), ATRIUM holds that categorisation can be construed as involving task switching, at the current stage it does not have any mechanism to produce switch costs. However, the modular architecture of the model and its gating network account have been verified to be successful in explaining contextual modulation of attention in category learning. The following simulation thus focuses not on establishing how to adapt ATRIUM's architecture to produce switch costs, but on demonstrating to what extent

the representational attention mechanism of ATRIUM can reflect the nature of task switching in category learning, because this is a necessary step to extend a model of task switching effects on category learning.

6.4.2 Category Structures

The category structures used here is the same as Erickson's (2008). Erickson (2008) used fixed-width rectangles with an internal vertical line as stimuli. The bottom and sides of the stimuli were white, while the internal line segments and the top of the rectangles were either cyan or magenta. Stimuli from categories A and B were drawn randomly with one of these colours and stimuli from categories C and D were drawn with the other. The stimuli varied on rectangle heights and line segment positions (see Fig 6-9). All parameter values for generating the stimuli are the same as those used in Erickson (2008).

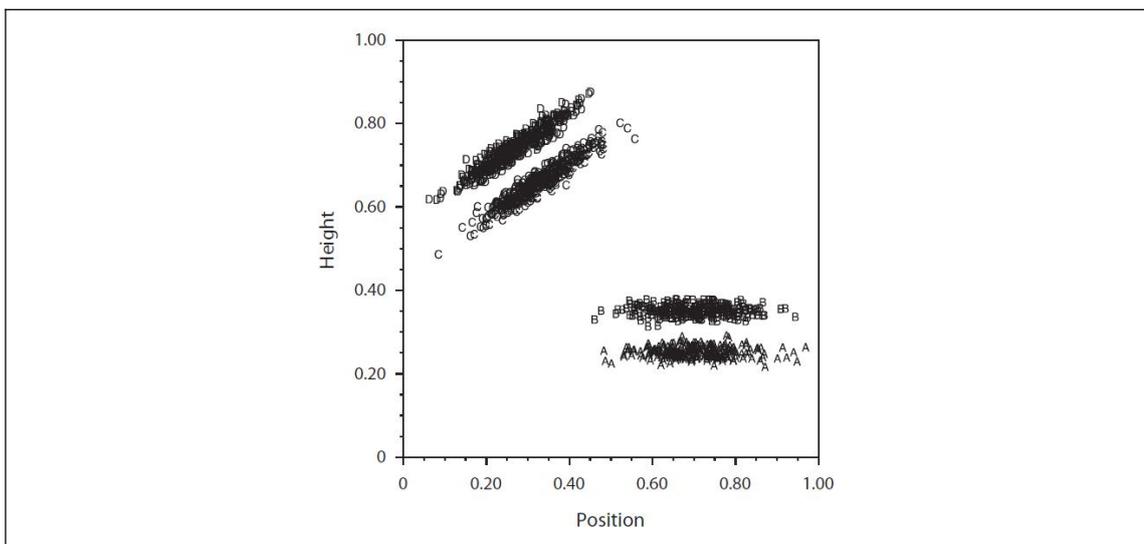


Fig 6-9. The category structures used in Erickson's (2008) Experiment. Each letter, A, B, C, or D, in the figure indicates an item of that category presented during the experiment. Categories A and B comprised the 1-D substructure, and categories C and D comprised the 2-D substructure.

6.4.3 A Modified Model

A schematic illustration of the modified model is shown in Fig 6-10. The modular architecture of ATRIUM and the gating mechanism are retained in the modified model. The equations of activations of all hidden nodes, response nodes and gating nodes are the same as the original ATRIUM, except that the rule nodes originally used in the 1D representation module

which implement a linear sigmoid activation rule are replaced by the exemplar nodes. In addition, the attention learning mechanism is omitted in the modified model. Instead, the strength of each dimension in this model is given by reasonable hand-set values (see Table 6-2). There are two gating nodes, each determines the extent to which the output of one internal module influences the overall responses. Learning in this model follows the standard error reduction mechanism.

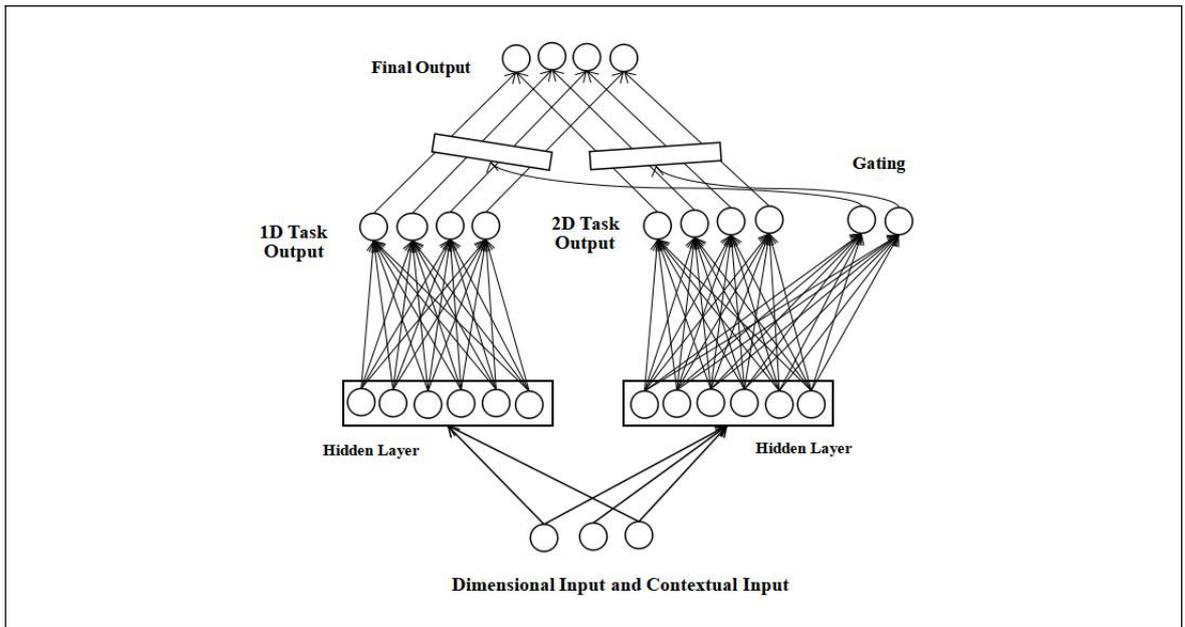


Fig 6-10. An illustrative representation of the modified version of ATRIUM model. Note that the modular architecture is designed towards the model of task switching effects in heterogeneous categorisation.

In the original ATRIUM model, error is defined as

$$E = \frac{1}{2} \sum_K (t_K - o_K)^2 \quad (6.6)$$

where t_K denotes a vector of target teacher values, and o_K denotes the vector of output nodes' activations. The error can also be understood in terms of the accuracy. If the accuracy of module m is defined as

$$A_m = \exp(-E_m) \quad (6.7)$$

The mean accuracy of the model, \bar{A} , can then be expressed

$$\bar{A} = \sum_m p_m A_m \quad (6.8)$$

so that the mean accuracy of the model can be regarded as a measure of learning performance of the model..

Table 6-2.
Parameter Functions and Settings Used in Section 6.4

| Parameter | Function | Settings |
|-------------|---------------------------------|----------|
| c | Specificity | 0.80 |
| ϕ | Decision consistency | 3.00 |
| y_g | Gate gain | 1.00 |
| β_g | Gate bias | 1.50 |
| η_{1D} | Learning rate of 1D module | 0.10 |
| η_{2D} | Learning rate of 2D module | 0.30 |
| η_g | Learning rate of gating network | 0.15 |
| a_c | Strength of context | 3.00 |

6.4.4 Method

In Erickson's (2008) experiment, participants were first trained on the 1D categories and 2D categories for 200 trials (4 blocks of 50 trials) and 550 trials (11 blocks of 50 trials), respectively. After then, they went through 4 blocks of 100 intermixed trials (i.e., 25 trials for each category). This simulation yielded the same procedure of training.

6.4.5 Results and Discussion

Fig 6-11 shows the change of mean accuracy of the model during training. As can be seen in the figure, the mean accuracy gradually increases, and so in the intermixed task switching phase (trials 751 onwards), the model performs nearly perfectly.

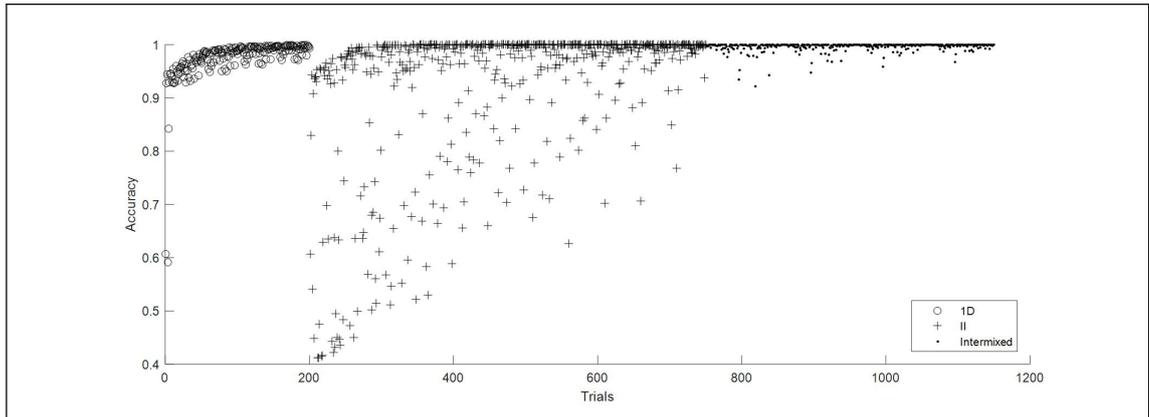


Fig 6-11. Change of mean accuracy of the modified model as a result of training. The first 200 trials involved a 1D rule-based structure. Trials 201 to 750 involved a 2D information-integration structure. Trials 751 onwards required switching between the two types of stimuli.

Fig 6-12 shows the gating weights of different modules, and in different contexts, as a result of training. Note that, again, in this simulation, we assume that the role of extra contextual cues is different from other stimulus dimensions. This is done by giving them greater attentional strength. As a result, as can be seen in Fig 6-12, the learned gating weights show different patterns in different contexts. For example, in context 1 (indicating 1D task), the influence from the 1D representational module is stronger than in context 2 (indicating II task), and vice versa. Nevertheless, this result reflects the principle of contextual modulation in category learning. It indicates that representational attention shares some properties with higher-level attentional control, such as task switching.

However, one thing that must be noticed is that the error-reduction mechanism used in the original ATRIUM model may not provide a good account for task switching. Although we have used a greater strength on the contextual cue, learning in the second task (i.e., II categories) cannot neglect the influence from the 1D task. In a preliminary study that uses the error reduction mechanism of ALCOVE to simulate the Wisconsin Card Sorting Task (see Appendix B), we have observed that the error reduction mechanism used in the original attention learning

model limits the capacity of the model for reconfiguration and shifting to a novel task set. The gating network of ATRIUM inherits the original learning algorithm, but this could be a fatal shortcoming for the current model to account for the Erickson data set. One possibility to solve this problem might be to introduce a hierarchical mixture-of-experts architecture (Jacobs et al., 1994). In a hierarchical mixture-of-experts architecture, a single task can be represented by a low-level mixture-of-experts substructure. A higher-level gating network coordinates the competition between these low-level mixture-of-experts structures. Learning in the higher-level gating network could not occur until the intermixed blocks. However, it must also be noticed that the purpose of extending the modular architecture of category learning should be not only to predict accuracy, but also to produce switch costs.

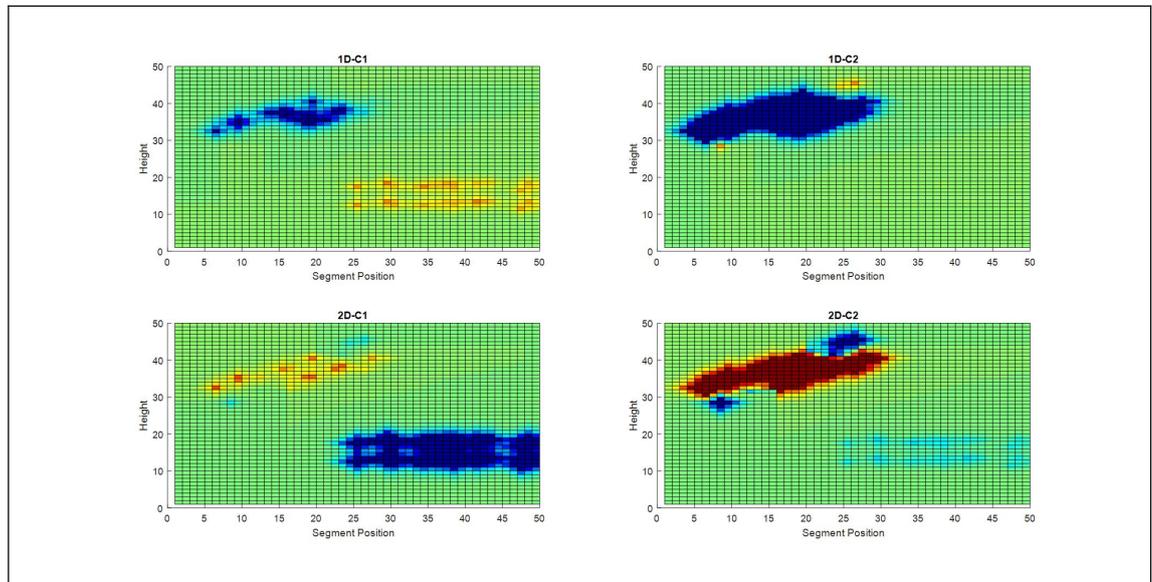


Fig 6-12. Weights in the gating network trained in Erickson’s experimental paradigm. The cool colours (e.g., blue) represent small weights whereas the warm colours (e.g., red) represent large weights. (1D = gating weights of one-dimensional representation module; 2D = gating weights of two-dimensional representation module; C1 = Context 1; C2 = Context 2)

In sum, the results of this simulation, together with the simulation in section 6.3, imply that although the gating network of ATRIUM does reflect the nature of attentional control, and in particular task switching, the representational attention account and the error reduction mechanism limit the model to capture the goal of theoretical integration in category learning.

6.5 Theoretical Implications

This chapter has illustrated the principles of a mixture-of-experts classifier through presentation of a reimplementation of the original version of ATRIUM for fitting a rule-based extrapolation phenomenon, and, then, applying it to II and hybrid categories with continuous-valued dimensions. While ATRIUM was able to perform the task moderately well, the Monte Carlo simulation suggested that ATRIUM is unable to keep a balance between learning II and hybrid categories. When II categories are controlled by the exemplar module, the same set of parameters appears to lead to strong trial-by-trial switching for hybrid categories. In contrast, when the rule module dominates hybrid categories, the same set of parameters appears to predict that the rule module will control responding on II categories, too.

6.5.1 Limitations of ATRIUM

Like COVIS, ATRIUM postulates that there are two qualitative distinct representational modules that mediate categorisation: a rule-based module and an exemplar-based module. There are three types of interactions between these modules. First, the gating network determines how much each expert network contributes to the output of the module. Unlike COVIS, the outputs of two modules are blended to produce the final output. As mentioned earlier, each gate's activation determines the weight of each module in the final output. Second, the performance of modules determine the output of the gating network, because learning in the gating network is driven by the overall error of each system on a trial-by-trial basis. Third, the gating network determines how much each module learns.

Indeed, the last two types of interactions would lead the ATRIUM model to produce a stimulus-specific representation. In contrast, as we can see in the simulation study of Chapter 5, the separate systems in COVIS compete to solve the entire categorisation task individually. In a sense, the implementation of COVIS can be considered a global learning adjustment or stimulus-independent classifier, whereas the mixture-of-experts model can be considered to be a

stimulus-dependent classifier. A problem with the stimulus-independent classifier is that it is unlikely to reproduce rule-like extrapolation phenomena, because for learning rule-plus-exception categories, COVIS would reduce to a single system model at the end of training. However, ATRIUM is inspired by heterogeneous, rule-plus-exception category representations whereas COVIS is inspired by homogeneous category representations. In heterogeneous category learning, it is highly likely that participants learn to apply a complex, disjunctive rule, such as ‘the stimuli with same values on two dimensions are exceptions, otherwise use 4.5 on height as the rule boundary’. In fact, this type of solution was found in Erickson and Kruschke (1998). However, the evidence is overwhelming that the optimal strategy for learning II categories has no verbal description (Ashby & O’Brien, 2007; Maddox et al., 2004a; 2004b; Maddox & Ing, 2005). II learning is mediated primarily by a procedural learning system, but not heterogeneous, rule-plus-exception representations (e.g., Ashby et al., 1998; Ashby & Maddox, 2005). Hence, there appears to be a conflict between ATRIUM and COVIS models on their predictions of representations of II categorisation.

In addition, the current architecture of ATRIUM does not allow the model to account for the production of complex rules. As it has been currently realised, ATRIUM does not fully implement a dual systems account, because the model does not have a rule selection process. Although it does select between rule and exemplar-based representations, it should also be able to test and select between rules, just like the explicit-learning system in COVIS. Lewandowsky and his colleagues (Sewell & Lewandowsky, 2012; Yang & Lewandowsky, 2004; Yang et al., 2006) have tried to extend the gating system. They applied the gating network to select between multiple rules in deterministic, heterogeneous category structures. But no attempt has been made to account for the interaction between qualitatively distinct representations.

Moreover, in the empirical study of multiple representations, a benchmark phenomenon is that human participants use the exemplar-based representation to learn II categories, and tend to use a rule-based strategy to solve hybrid categorisation tasks. The simulation studies have shown that the COVIS model is able to balance between learning these categories, but ATRIUM seems unable to account for these phenomena. However, another empirical finding is that trial-by-trial switching is possible when participants are provided with extra switching signals (e.g., Crossley et al., 2017; Erickson, 2008). Erickson (2008) suggested that switching between different

representations in category learning is a special case of task switching (see also Crossley et al., 2017), because, although trial-by-trial switching is possible, it is not easy. Thus, one possibility is that control of multiple representations in human categorisation involves both mechanisms. In a modular architecture of category learning, the stimulus-dependent representation mechanism can be considered a bottom-up mechanism, whereas the stimulus-independent mechanism may involve the top-down process. This issue will be discussed in more detail with respect to this dual component control mechanism in Chapter 7.

6.5.2 The Organisation of Multiple Internal Representations

According to the mixture-of-experts model, each expert module can be regarded as a set of learned stimulus-response associations. While the exemplar-based system is mediated by associations between each individual object and category responses, the rule-based system is mediated by abstract rules. In this sense, the function of the gating network throughout the whole training process is to develop the organisation of separate learned stimulus-response associations. Thus, while completing the training phase, the relationships between each exemplar and separate experts is constructed. The environment changes. The changing environment may need a mechanism for the control of action selection that involves the flexible interplay between learned action selection and cognitive control. Although, the mixture-of-experts approach is able to account for some properties of bottom-up control in the organisation of learned action selection, there is no component related directly to top-down control.

Given that ATRIUM shares some of these properties of organisation of learned action selection, it may appear plausible to expect that applying some principles of action control theory could allow the extension of the mixture-of-experts approach to account for interesting phenomena. According to a well-known supervisory system theory, beyond the organisation of learned behaviour, there exists some processes occasionally invoked in regulating and generating responses, I assume that this cognitive control system can also initiate or override category learning when necessary. Erickson (2008) also argued that the control mechanism mediating task switching in category learning, at a conceptual level, ‘seems to share a number of properties with

the supervisory attentional system described by Norman and Shallice (1986)' (p. 750). However, at the same time, Erickson (2008) acknowledged that there is no appropriate component accounting for this top-down control.

Cooper and Shallice (2000), in their description of a model of learned action selection control, argued that: 'Objects with highly active representations tend to trigger relevant schemas more than equivalent objects with less active representations, and objects whose representations are inhibited below the resting activation inhibit relevant schemas.' (pp. 311). They further argued that the operation of action selection also involves the objects' activation-based argument triggering action selection. In fact, the function of the argument triggering an action is similar to the concept of the gating network. However, the original supervisory system theory did not account for learning. According to the Norman-Shallice model, the bottom-up control of learned action selection is determined by a contention scheduling (CS) system. The CS system reflects the function of memory (Bobrow & Norman, 1975). Unfortunately, Shallice and his colleagues have yet to clarify where these memories come from, but they did argue that the supervisory system can come into play in action control when contention scheduling is unable to produce an optimal response.

A model accounting for the organisation of learned categorisation behaviours must be able to implement learning of different categories and apply learned complex rules driven by exemplar-similarity. The mixture-of-experts model has strengths in accounting for these phenomena. At the same time, the gating network of this model has been shown to share many properties with the bottom-up control of sequential action selection. However, the gating network and error reduction might not be enough to account for situations, such as continuous-valued, probabilistic categorisation and system switching, which appear to require the supervisory system coming into play. These issues will be reconsidered in Chapter 7.

6.5.3 Implications for Task Switching

Although the ATRIUM model has some limitations, the modular architecture remains meaningful for task switching. The modular architecture is necessary to modelling task

switching. Most network models of task switching consist of multiple internal representation modules. These modules compete with and interactively activate one another to dominate the control of response generation. The top-down control mechanism mediates the processes by inhibiting responses from inappropriate modules and exciting the appropriate module. But, the traditional task switching models have not included exemplar-based representations. On the other hand, though the gating network provides a good locus for extending the bottom-up control mechanism in task switching, the modular architecture of category learning has not included any top-down control mechanism. Therefore, for modelling the representation of heterogeneous categorisation, a plausible solution might be to combine the two approaches into one modular network.

Chapter 7.

Attention to Multiple Representations: The Perspective from Supervisory System Theory

7.1 Introduction

There is a lot of evidence in recent years showing that there is a subtle correlation between attentional control and category learning. As mentioned earlier, evidence from Erickson (2008) revealed that overall performance of category learning seems to rely on the functioning of the attentional control mechanism (Engle & Kane, 2005). Further research has recently confirmed the associations between category learning and the attentional control mechanism (e.g., Craig & Lewandowsky, 2012; Kalish et al., 2017; Lewandowsky, 2011; Sewell & Lewandowsky, 2012). Many studies have, in addition, found that both ERB and II category learning are impaired when attentional control mechanism is not fully available (e.g., Miles et al., 2014; Schnyer et al., 2009; Maddox et al., 2010b). Evidence from both sides, again, opposes the multiple systems theory of category learning.

Despite abundant evidence for the importance of attentional control in category learning, no attempt has yet been made to develop a network model that incorporates an attentional control mechanism in categorisation. This may be because, on one hand, the traditional experimental paradigms and the dual category learning systems theory limit the thinking of combining cognitive control and categorisation. The primary focus in conventional cognitive control is on establishing a mechanism consisting of functions that mediate the selection of appropriate actions. These control processes include regulation and monitoring of multiple representations.

Though, categorisation, too, involves the process of selection among multiple representations, no empirical data has been reported to reveal the properties of this selection process. Instead, it was simply asserted that multiple independent representational systems compete to determine the final output. However, recent research on task switching in categorisation provides a new opportunity for us to further understand the role of attentional control in categorisation (e.g., Helie, 2017).

The failure to develop such a combined model may be due to the current lack of a computational basis that instantiates attentional control in object categorisation. Erickson and Kruschke (1998) suggest that in addition to dimensional attention, there is a kind of higher-level, representational attention. In the multiple representations context, representational attention inhibits inappropriate representational influences, and, thus, easily instantiates switching between representations. Dimensional attention is necessary to reflect the change of dimensional relevance in learning a homogeneous structure, but seems unrelated to performance on heterogeneous category learning. As mentioned in Chapter 5, representational attention reflects the influence from segregated representations based on various dimensional attention distribution patterns. Erickson (2008) suggested that representational attention, at a conceptual level, shares a number of properties with the attentional control mechanism (see also Craig & Lewandowsky, 2012; Kalish et al., 2017; Lewandowsky, 2011; Lewandowsky et al., 2012; Newell et al., 2010).

However, though, it has been verified to have many advantages over the original feed-forward, attention learning network, the mixture-of-experts approach, potentially, remains problematic. First, the gating network in ATRIUM only implements an exemplar-based, bottom-up processes, but ATRIUM does not include the requisite top-down control process. Second, according to ATRIUM, the rule module represents a kind of attention allocated to a single dimension only, whereas, at the same time, the exemplar module still learns to adjust dimensional attention strengths to dimensions. In a sense, the attention distribution is divided into two segregated processing pathways. Of course, this problematic implementation is influenced by the dual learning systems account. Third, in the implementation of ATRIUM, context is considered no different from the other features of the stimulus. This assumption is inconsistent with what is found in the recent literature where it has been shown that contextual

input can modulate attention to dimensions, but other dimensional input cannot (e.g., Ashby & Crossley, 2010; Crossley et al., 2017; George & Kruschke, 2012; Helie, 2017).

An alternative solution is to introduce the perspective from attentional control theory into the attention learning framework. This idea is motivated by two things. First, attentional control is essential for task switching. Erickson (2008) suggested that control of task switching in categorisation shares some properties of attentional control. Second, interestingly, research on automatisisation in categorisation has revealed that, analogous to the idea that control of non-automatic and automatic action selection is qualitatively distinct, there are dissociations between pre-automatic and automatic categorisation (Ashby & Crossley, 2012; Norman & Shallice, 1986; Waldschmidt & Ashby, 2011). This is consistent with the attentional control theory (see below). Meanwhile, a growing body of research has also argued that there should be an integrated control mechanism in pre-automatic categorisation.

This Chapter therefore introduces an influential attentional control theory – the Supervisory System theory, and some of its instantiated models – and discusses how it may be combined with the mechanisms of categorisation.

7.2 Supervisory System: A Theory of Cognitive Control

7.2.1 Supervisory System

Norman and Shallice (1986) proposed an informal model of the control of action, assuming that our daily behaviours are controlled by separate systems. Control of automatic behaviours requires minimal attentional resources, and is mediated by a low-level system referred to as *Contention Scheduling (CS)*. Operation of CS is autonomous and held to rely on the trade-off between well-established internal representations and environmentally triggered, bottom-up affordances. For example, when viewing a German Shepard, people immediately respond *dog* rather than *wolf*, even though such a response might require integrating perceptual information about the shape and size of the ears, the length, coarseness, colour of the hair and many other perceptual features. According to the standard cognitive psychology literature, such behaviours

can be performed in parallel and with the least conscious awareness when the representation has already been well-established in the mind (Karmiloff-Smith, 1986; Schneider & Shiffrin, 1977).

In contrast, dealing with novel, changing or conflicting situations requires the involvement of extra attentional resources. Control of non-automatic behaviours thus requires a higher-level, deliberative control system. This second control system should be able to modulate or override the operation of the CS system when required or desired (e.g. Luria, 1966; Fuster, 1989; Dehaene & Changeux, 1997; Miller & Cohen, 2001; Shallice, 2006). For example, in the task switching context the environment keeps changing. It is reasonable for there to be a mechanism to avoid conflict between an experienced and a forthcoming behaviour. Thus, it is assumed that control of non-automatic behaviours requires executive functions, such as active maintenance of relevant working memory representations and set-shifting (Shallice et al., 2008). These executive functions are operated by the *Supervisory System* (SS).

The SS can occasionally modulate the CS in order to achieve multiple tasks. The representation a task is referred to as a schema (Norman & Shallice, 1986). A schema in everyday life, such as preparing a cup of instant coffee, could be further decomposed into multiple internal sub-representations — sub-schemas. These sub-schemas are coherent because they achieve a schema's sub-tasks (e.g., sweeten the coffee), and they may be used in different tasks (e.g., preparing a cup of tea). The success of achieving a task is determined by the activation and selection in this schema/sub-schema architecture from moment to moment. According to Norman-Shallice's (1986) informal model (see Fig 7-1), the SS can organise, coordinate and monitor the selection of schemas at any moment when CS alone is insufficient to achieve the task, acting as a general-purpose planning component. This is necessary because the operation of CS is dominated by the familiarity and frequency of application of tasks (Shallice, 2006) – it is unable to resolve conflicts caused by the changing environment. This structure is useful for explaining many effects in task switching.

The Supervisory System is strongly associated with frontal cortex. Multiple earlier neuropsychological studies have shown that lesions in different parts of frontal cortex could produce different fashions of impairment on performing tasks that require switching and other supervisory functions, such as the Wisconsin Card Sorting Test (Milner, 1963) (Stuss et al., 2000; see Shallice et al., 2008 for review). In addition, in cognitive neuroscience studies, many regions

in frontal cortex are, again, found associated with functions like task switching, set-shifting and response reversal (Buchsbaum et al., 2005; Derrfuss et al., 2005). In addition, note that the attentional control theory also suggests a decreasing need for attention with practice. This prediction is consistent with neuroimaging studies that show that the activation of regions of frontal cortex are greater during the early stages of sequence learning (e.g., Jueptner et al., 1997; Passingham et al., 2005), and neuropsychological studies that show that motor and category learning are impaired by frontal dysfunction (Richer et al., 1999; Schnyer et al., 2009; Maddox et al., 2010b). In sum, according to the attentional control theory, control of task switching is associated with functioning of frontal cortex.

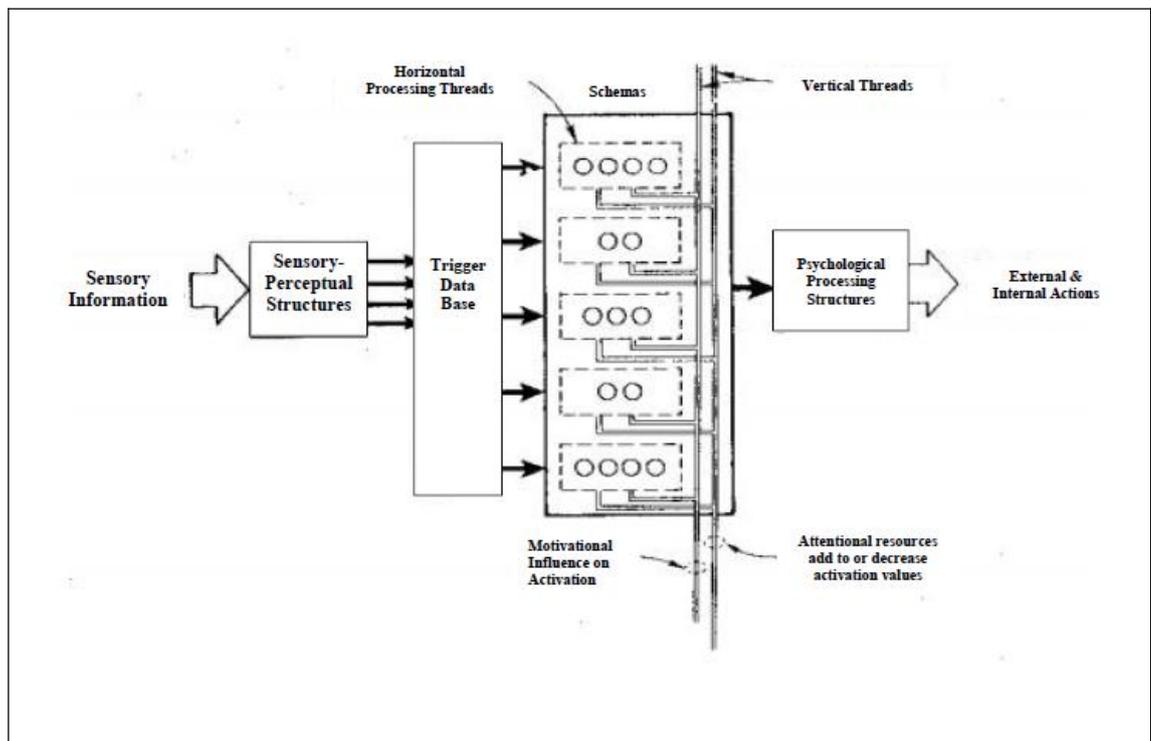


Fig 7-1. The schematic illustration of cognitive control according to Norman-Shallice model.

7.2.2 The Supervisory System and Task Switching

In the 1990s, empirical studies in the task switching paradigm showed that if stimuli require different responses depending on which task is operative, then trials where the task switches

from the previous trial (switch trials) are slower than those where the task repeats (stay trials) — the difference between the two reaction times is known as switch cost (Allport et al., 1994; Rogers & Monsell, 1995; see Kiesel et al., 2010 for a review). Two types of explanation have been proposed. Rogers and Monsell (1995) proposed that there is a top-down control mechanism of task reconfiguration, because in their studies they found that the size of the switch costs can be decreased with an increase in the time available in which the participant can prepare for the specific task (Meiran, 1996; Rogers & Monsell, 1995). This approach provided a measure of the time to switch the dimension of the stimulus that should control the behaviour. However, Allport and colleagues (1994) proposed the control mechanism known as backward inhibition, because Allport et al. (1994) found that the switch costs were less when switching is into a more difficult task than into an easier task. The backward inhibition account suggests that the switch costs are due to the need to inhibit the previously active task representation.

Many neuroimaging studies have found many regions in frontal cortex that show some degree of association with task switching performance (Buchsbaum et al., 2005; Derrfuss et al., 2005). Moreover, a large number of task switching studies have been carried out in neurological patients (Aron et al., 2004; Mayr et al., 2006; Shallice et al., 2008). For instance, Aron et al. (2004) found that both patients with left frontal lesions and right frontal lesions showed significantly larger switch costs than controls. The correlation between frontal cortex and task switching suggests that the Supervisory System has a role to play.

From the perspective of Supervisory System theory, the task reconfiguration account and backward inhibition account are not necessarily in conflict. In a task switching model that embodies Supervisory System control, these two types of explanation have been represented at different levels of a modular architecture of interactive activation type (Gilbert & Shallice, 2002). On such a model, though, the switch cost as measured from reaction times, no longer corresponds quantitatively to the time that the internal process of task reconfiguration takes. The next section introduces this interactive activation and competition network model.

7.3 A Modular Model of Task Switching

7.3.1 The Gilbert-Shallice (2002) Model of Task Switching

How can we computationally instantiate the Supervisory System theory in the task switching context? To date, a variety of network models of task switching have been proposed, most of which are drawn on the principles of interactive activation and competition (IAC) (e.g., Botvinick et al., 2001; Brown et al., 2007; Cooper & Shallice, 2000; Gilbert & Shallice, 2002). Gilbert and Shallice (2002) proposed an IAC network model of the Stroop task (see Fig 7-2), instantiating some principles of the supervisory system theory.

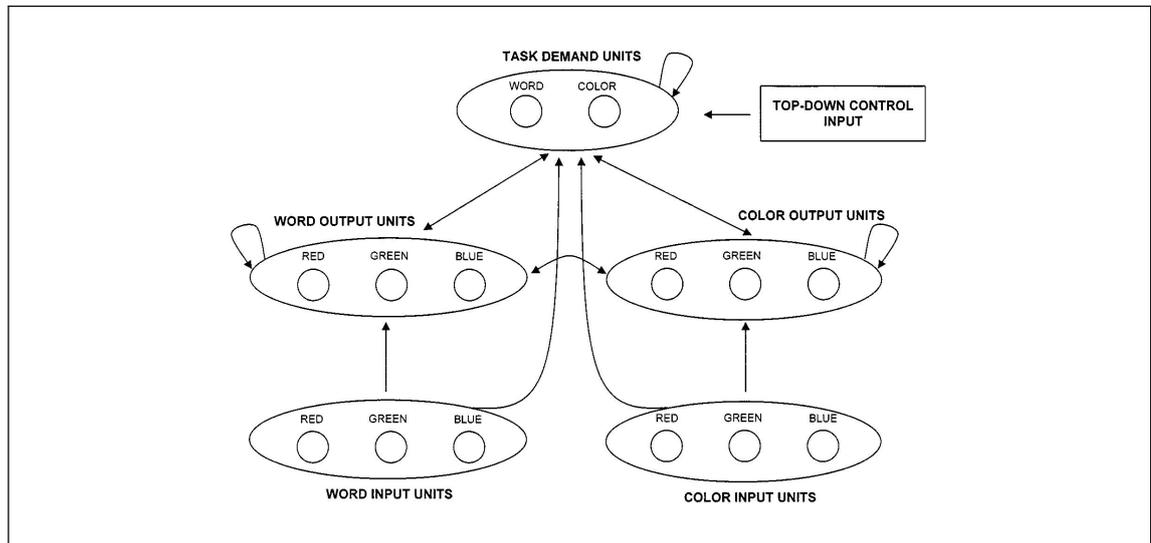


Fig 7-2. An Illustrative representation of Gilbert-Shallice (2002) Model. Note that this models switching between word reading and colour naming tasks.

In the Gilbert-Shallice (2002) task switching model, the input alternatives across each stimulus' dimension and each possible response are represented as single units. Task switching is implemented on the basis of two segregated modules (see Fig 7-2). The weights of the associative links from stimulus input to response units are greater in the word reading module than the colour naming module, reflecting greater experience with this task. Two 'task demand' units provide control of the current task. According to Gilbert and Shallice (2002), the task demand units reflect activity in prefrontal cortex (see also Miller & Cohen, 2001). The task demand units can excite all response units for their respective task and inhibit the alternative task

response units. Lateral inhibition is used between units at the same level. Thus, for example, lateral inhibition between the two task demand units ensures that they compete to become the most highly active, with activation of either suppressing activation in the others.

In the absence of an attentional control mechanism, the network produces responses from the stronger, word reading module. To perform task switching, therefore, a top-down control input is applied to the relevant task-demand unit on each trial. As more biasing is required to perform the weaker task than the stronger one, the top-down control input, for example, is stronger for colour naming than word reading. On a typical trial, stimulus input units are activated for each module (e.g., the green colour unit and the red word unit representing the word 'red' incongruously displayed in green), with activation propagating through both modules. Meanwhile, a single task-demand unit is activated by top-down control excitation. This biases processing in favour of the current task, while suppressing processing in the competing task. In the model, interference from the previous trial is produced due to the residual activation of the task-demand units from that trial. Repeat trials, thus, are facilitated, as the relevant task-demand unit remains highly active. Interference occurs on switch trials, as residual task-demand activation now facilitates processing for the competing task, and a greater period of processing is required for top-down control to re-activate the expected task-demand unit against previous task interference. Switch costs are then produced.

The Gilbert-Shallice (2002) network model implements the assumption that SS biases the CS by providing a top-down control input favouring the appropriate task (schema), thus effectively over-riding the dominant response from CS alone. It does not, however, implement the decreasing need for attention. Elements of the model can be easily mapped onto the attentional control theory. The internal representation of a task corresponds to the task demand units, the bottom-up processing corresponds to the associations between the stimulus input units and the task demand units, and the influence from SS corresponds to the top-down control input into the task demand units. Therefore, in theoretical and practical terms, the model provides a foothold for implementing task switching in categorisation.

7.3.2 Illustrative Example: Modelling RT Costs in Task Switching

To illustrate how the Gilbert-Shallice (2002) model works, a reimplementations of the model was developed and applied to simulate task switching based on few Stroop stimuli.

7.3.2.1 Method

The stimulated stimuli were colour words 'Red', 'Blue' and 'Green' presented in same (congruent) or different (incongruent) colour inks. The model's parameter settings were the same as the standard settings in the original model (see Table 7-1). The task switching paradigm was simulated with a run length of four before each switch of task. Thus, the model performed four trials of word reading, followed by four trials of colour naming, and so on. This simulated the experimental procedure designed by Rogers and Monsell (1995) Experiment 6. For illustrative purposes, 36 trials were used.

7.3.2.2. Results

The mean reaction times of each trial (see Fig 7-3A) and RT costs on stay and switch trials (see Fig 7-3B) were recorded. As can be seen from Fig 3, the model produces switch costs. That is, switch trials have longer response times than stay trials. In addition, Fig 7-3A shows a decrease of RT costs following each switch trial. This is because the residual activation from the previous trial in the task demand units allows the task active in the previous trial to continue to influence the activation on the output level. In a switch trial, it requires more cycles to activate the task, whereas, in a stay trial, the number of cycles required for reactivation is less.

Table 7-1

Model parameter settings for simulating task switching

| | Parameter | Settings |
|----|---|----------|
| 1 | Maximum activation level | 1.00 |
| 2 | Minimum activation level | -1.00 |
| 3 | Response threshold | 0.15 |
| 4 | Step size | 0.0015 |
| 5 | Squashing of task demand units | 0.80 |
| 6 | Noise | 0.006 |
| 7 | Output units bias | -6.00 |
| 8 | Task demand units bias | -4.00 |
| 9 | Stimulus input strengths (Word) | 3.00 |
| 10 | Stimulus input strengths (Colour) | 1.90 |
| 11 | Top-down control input strength (Word) | 6.00 |
| 12 | Top-down control input strength (Colour) | 15.00 |
| 13 | Lateral inhibition (within word/ colour module outputs and between task demand units) | -2.00 |
| 14 | Between modules inhibition | 2.00 |
| 15 | Between modules exhibition | 2.00 |
| 16 | Output-task demand connection strengths | 1.00 |
| 17 | Task demand-output connection strengths | 2.50 |
| 18 | Learning rate (adjusting connection weights between stimulus input and task demand units) | 1.00 |

Note: Parameters 16 and 17 determine both positive and negative connection strengths between units.

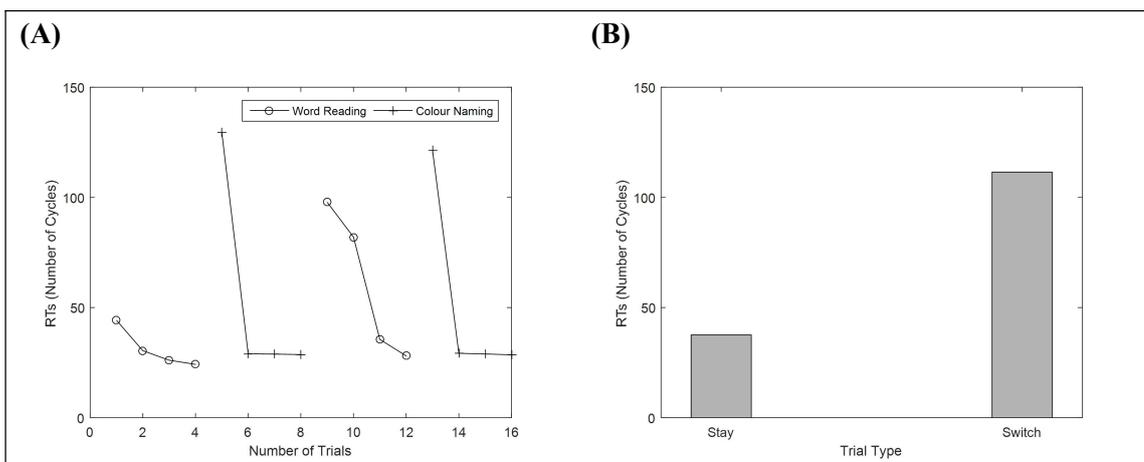


Fig 7-3. (A). Performance of the model in 16 simulated trials; (B). The simulated mean reaction time costs represented by number of cycles in switching between word reading and colour naming tasks. These results are based on 100 times of simulation.

7.3.3 Implications for Multiple Representations Theory

The principle of the Gilbert-Shallice (2002) model is simple. Now, consider that if we regard the Stroop stimuli as exemplars, and word reading and colour naming tasks as different categorisation tasks. The word reading task requires attentional focus on the feature of alphabet, whereas the colour naming requires attentional focus on ink colour. This model has, effectively, implemented a heterogeneous categorisation paradigm. There is only one problem in this model architecture: the complexity of the modular feed-forward network is underestimated. The model, thus, is unable to be directly applied to modelling the representation of heterogeneous categorisation tasks.

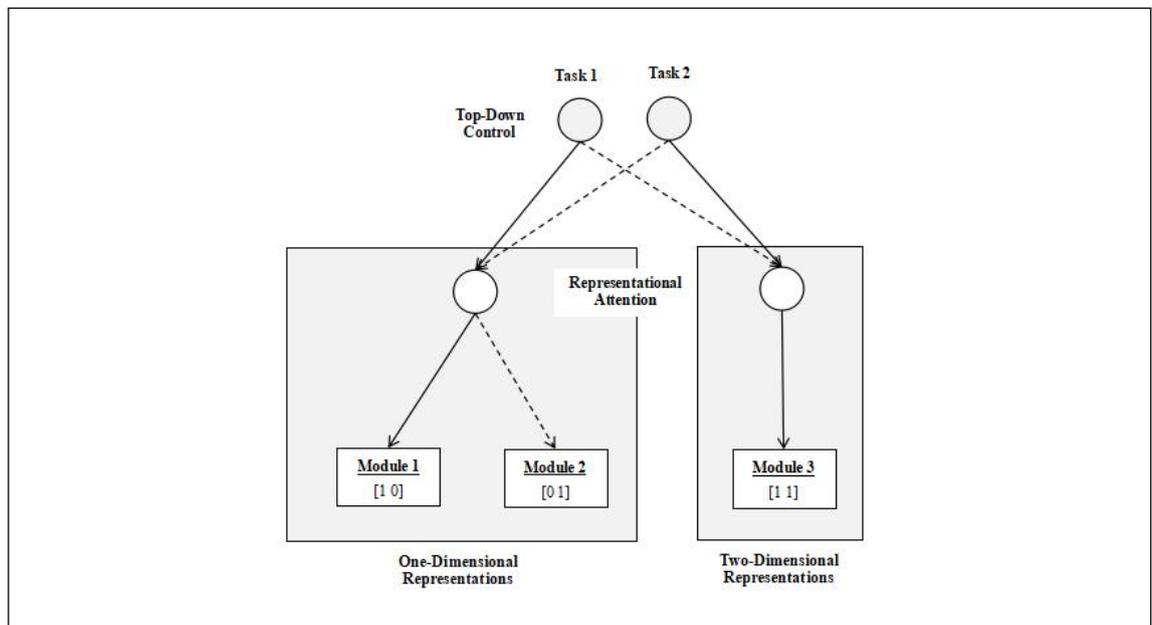


Fig 7-4. A schematic illustration of the modular architecture of heterogeneous category representation of two-dimensional stimuli. In the model, internal representation is determined by the focus of attention on each dimension. Modules 1 and 2 represent the representation determined by single dimensional attention, and Module 3 is determined by multidimensional attention. Each internal representation module is proceeded in parallel. Activation of task demand units inhibit and excite different internal representations. Dashed lines represent inhibitory connections, whereas solid lines represent excitatory connections.

However, it may not be so difficult to solve this problem. The easiest way is to combine the model architecture with the multiple representations theory. Fig 7-4 shows the schematic representation of a modular architecture of categorisation. Each of the internal representation modules in this network is replaced by an exemplar-based association network. Activation of exemplars in these modules is determined by the focus of attention to each stimulus- dimension.

Thus, for a stimulus varying on two continuous-valued dimensions, only three alternative modules are included (two 1D representations and one 2D representation). It is assumed that, in homogeneous category learning, the model learns to excite the appropriate internal representation module and inhibit others, whereas, in heterogeneous category learning, activation of different task demand units excite and inhibit different internal representation modules in response to changes of contexts. Therefore, this combined modular network provides a platform for considering task switching control in categorisation.

7.4 The Hyperdirect Pathway Hypothesis

How does frontal cortex affect the organisation of multiple internal representations in categorisation? Research in cognitive neuroscience has confirmed that a critical neural structure in category learning is the basal ganglia (BG). The BG affects human behaviours via three pathways connecting the cortex to the thalamus: the direct, indirect (Alexander & Crutcher, 1990), and hyperdirect (Nambu et al., 2000) pathways. The direct pathway contributes to action selection by reducing the inhibition from the GPi/SNr on the thalamus. As can be seen in Fig 7-5, inhibitory projections from the GPi/SNr to the thalamus, exert a tonic inhibition that keeps all potential behaviours suppressed. When an appropriate behaviour is identified, this tonic inhibition is reduced for the selected action, which then is executed. Therefore, the direct pathway is also termed the ‘go’ pathway because it results in the release of a movement. In contrast, the indirect pathways is involved in suppressing actions by increasing the inhibition from the GPi/SNr on the thalamus. This pathway is also called the ‘stop’ pathway (Frank, 2005). Unlike the direct and indirect pathways, the hyperdirect pathway passes through the subthalamic nucleus (STN) rather than the striatum. The projections from cortex to STN are more diffuse than the projections from cortex to striatum, and as a result the hyperdirect pathway’s effects on the thalamus are less specific to particular stimuli and responses than the other two pathways. Importantly, it has been suggested that the frontal cortex’s control through the hyperdirect pathway is able to stop behaviours that have already begun execution and prevent premature responding (Aron & Poldrack, 2006; Frank, 2006; see Seger, 2008 for review).

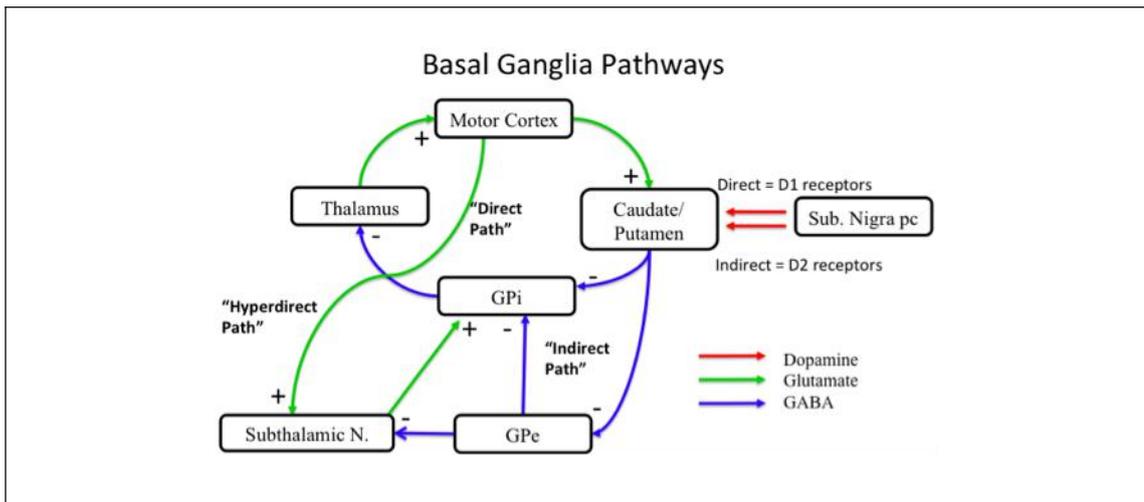


Fig 7-5. A simplified schematic illustration of the direct, indirect, and hyperdirect pathways' connectivity.

Ashby and Crossley (2010) proposed that one possibility is that the control from frontal cortex influences task switching in categorisation through the hyperdirect pathway. The hyperdirect pathway begins with direct excitatory projections from frontal cortex to the STN, which in turn can affect communication between the BG and the cortex. Specifically, the STN sends excitatory projections directly to the internal segment of the globus pallidus (GPi). This extra excitatory input to the GPi tends to offset inhibitory input from the striatum, making it more difficult for striatal activity to affect the cortex. Since responses from the procedural learning system are generated in the BG, decreased communication between the BG and cortex blocks the execution of responses from the procedural learning system. As a result, this pathway could permit (by reducing STN activity), or prevent (by increasing STN activity), signals coming from the striatum from influencing premotor cortical regions.

The above idea stems from research on the stop-signal task. On a typical stop-signal trial, participants initiate a motor response as quickly as possible when a cue is presented. On some trials, however, a second cue is presented soon after the first and in these cases participants are required to inhibit their response. A variety of evidence implicates the STN in this task. A popular model is that the second cue generates a "stop signal" in PFC that is rapidly transmitted to the GPi via the hyperdirect pathway, where it cancels out the 'go signal' being sent through the striatum. When the explicit learning system is in control, a similar stop signal may be used to inhibit a potentially competing response signal generated by the procedural learning system. One possibility may be that the primary role of the VMPFC is to control the hyperdirect pathway.

Schnyer et al. (2009) found that VMPFC patients were impaired in II tasks because they were more likely to use explicit rules, it is therefore reasonable to infer that damage to VMPFC would disrupt the normal transition from RB to II strategies. In other words, the default state (i.e., explicit learning domination) of the hyperdirect projection from PFC to the STN may be 'on' and one role of the VMPFC may be to switch this excitatory projection off.

The frontal cortex's ability to inhibit communication between the BG and the cortex via the hyperdirect pathway could be the mechanism by which explicit strategies dominate the hybrid category learning task, because it provides a mechanism via which the PFC can inhibit a response selected by the striatum, but it does not allow the striatum to inhibit a response selected by the PFC. In addition, this hypothesis could also account for the success of Erickson's (2008) participants on task switching of heterogeneous category learning (and also those of Crossley et al., 2017; Helie, 2017). The extra cues introduced in task switching could be sufficient to inform participants when to turn this signal on and off. Moreover, note that this account is based on the neuroscience data suggesting that the inhibition between systems is at the output stage. The hyperdirect pathway has no direct effect on processing within the striatum. Thus, importantly, control through the hyperdirect pathway hypothesis predicts that when the explicit learning system is in control, the procedural learning system operates normally but is blocked from motor output.

7.5 Toward a New Modular Architecture of Human Categorisation

7.5.1 Modular Architecture of Category Learning and Task Switching

In our daily life, we make thousands of categorisation decisions effortlessly per day. Many of them require selection among multiple internal representations, including some situations of task switching. Indeed, task switching based on category knowledge is an essential capability in human cognition. Although, behavioural evidence suggests that, in the homogeneous category learning context, the capacity of cognitive control does not predict strategy use, overall

performance of individuals with higher working memory capacity is better than those with low working memory capacity (e.g., Craig & Lewandowsky, 2012; Kalish et al., 2017; Lewandowsky, 2013). Effects of attentional control on performing heterogeneous categorisation are robust (Erickson, 2008; Swell & Lewandowsky, 2012). But, the problem is that most of the traditional multiple representations accounts of category learning are weak in the consideration of cognitive control.

The Supervisory System theory argues that a prefrontally-based, higher-level control system can modulate behaviours at any necessary point. As a system operating without (or with less) consciousness, the CS system should be able to generate candidate responses based on the environmental input. In other words, the CS system contains all the learned specific mappings between stimuli and responses, whereas the SS comes into play once any conflict or change appears. How can these points be mapped onto the control mechanisms of heterogeneous categorisation? According to the dominant neurobiological view of categorisation, categorisation is mediated by several cortico-basal ganglia-cortical loops. The BG is critical for category learning, whereas PFC mediates task switching. It is, thus, assumed that control of heterogeneous category learning may be based on two control components: a prefrontal control component, corresponding to a task switching mechanism, and a striatal control component that suits the need for establishing the stimulus-response mappings.

The representational attention determined by the gating network is suitable for task switching, because it predicts stimulus-dependent representation, and, more importantly, the gating network is trainable (e.g., Erickson, 2008; Kruschke, 2001; Sewell & Lewandowsky, 2011). Cooper and Shallice (2000; 2006) argued that a task (schema) can be divided into several subtasks (subschemas), and operation in this structure can be determined by the relationships between the schemas and influences from the current input stimulus. Likewise, in categorisation, people learn to exhibit the internal representations that optimise accuracy, and suppress the inappropriate ones (e.g., Lewandowsky et al., 2006; Yang & Lewandowsky, 2003; 2004). In other words, the gating network implements selection of appropriate subtask representations on the basis of learned associations between exemplars and gate units. However, unlike the original idea of ATRIUM, the evidence shows that trial-by-trial switching between ERB and II categorisation tasks does not take place easily. ATRIUM is inspired by a simple

rule-plus-exception task which needs switching between local strategies, but considerable evidence confirms that learning II categories requires a global change of strategies. The use of stimulus-dependent representations is problematic. As was shown in Chapter 5, ATRIUM sometimes even displays trial-by-trial switching in learning II categories. Furthermore, in addition to predicting accuracy performance, an integrative model should also be able to predict switch costs on reaction time (RTs), but the representational attention account was not motivated to fit RT data (Erickson, 2008). Although the representational attention mechanism, theoretically, can easily implement switching between tasks/subtasks, it still needs to be scrutinised and, of course, modelled.

Newell, Dunn and Kalish.'s (2011) reviewed the evidence from both behavioural and neuroimaging studies that supports COVIS, and concluded that the dual memory systems theory is fundamentally problematic. They argued that researchers in categorisation should escape the shackles of the dual memory systems explanations. A growing body of research reveals that a non-dualism explanation is also well-suited for human categorisation (e.g., Kalish et al., 2017; Miles et al., 2011; Newell et al., 2010; Sewell & Lewandowsky, 2012). In addition, lesion studies of prefrontal patients (Schnyer et al., 2009) and behavioural studies (Miles et al. 2011) taxing prefrontal functioning also revealed that the prefrontal control component is important for categorisation. Note that these arguments do not attack the assumption that the striatal control component is important. Here, it is supposed that the two control components are not implemented in COVIS fashion.

Given that the traditional modular networks of category learning are not enough, how can we computationally instantiate this hypothesis? For reasons of simplicity, an independent control mechanism, in addition to categorisation systems which produce separate internal representations, might be assumed, as Crossley et al. (2017) argued that trial-by-trial switching between ERB and II tasks is a special case of task switching. Following this idea, Sebastian Helie (2017), in his recent research, borrowed two factors (i.e., practice and preparation time) from the traditional task switching paradigm. He found that though preparation time and practice separately only had small effects on task switching, combining them significantly reduced the switch costs on reaction time. Helie (2017), therefore, suggested that the preparation time may have been initially too short, but as individuals became more proficient at task switching it may

have become sufficient. This result therefore also indicates that the control mechanism can be independent of the internal representation modules.

7.5.2 The Dual Control Component Account

It has been argued above that establishing a complete account of task switching in categorisation may reasonably need the combination of multiple internal representation modules and an independent task switching mechanism. Based upon the existing modular architecture of category learning and the Supervisory System theory of task switching, I propose a combined computational account of attentional control. The schematic illustration of the model is shown in Fig 7-6. As can be seen in Fig 7-6, the top-down control process (i.e., denoted as ‘task demand’), which is driven by contextual inputs, can determine which dimension to focus, and inhibit and excite modular outputs. The association network of each internal module that reflects the traditional exemplar-to-category response mappings implements the dimensional attention determined by task demand. The design is, in particular, suitable for task switching. In the task switching paradigm, the dimensional attention distribution associated with different tasks has already been learned. All the attentional control needs to do is, therefore, to select between task settings. In other words, in the process, the top-down control process defines where to focus, but not what to respond when the behaviour has become well-learned. This is consistent with the assumption that influence from the SS decreases during learning.

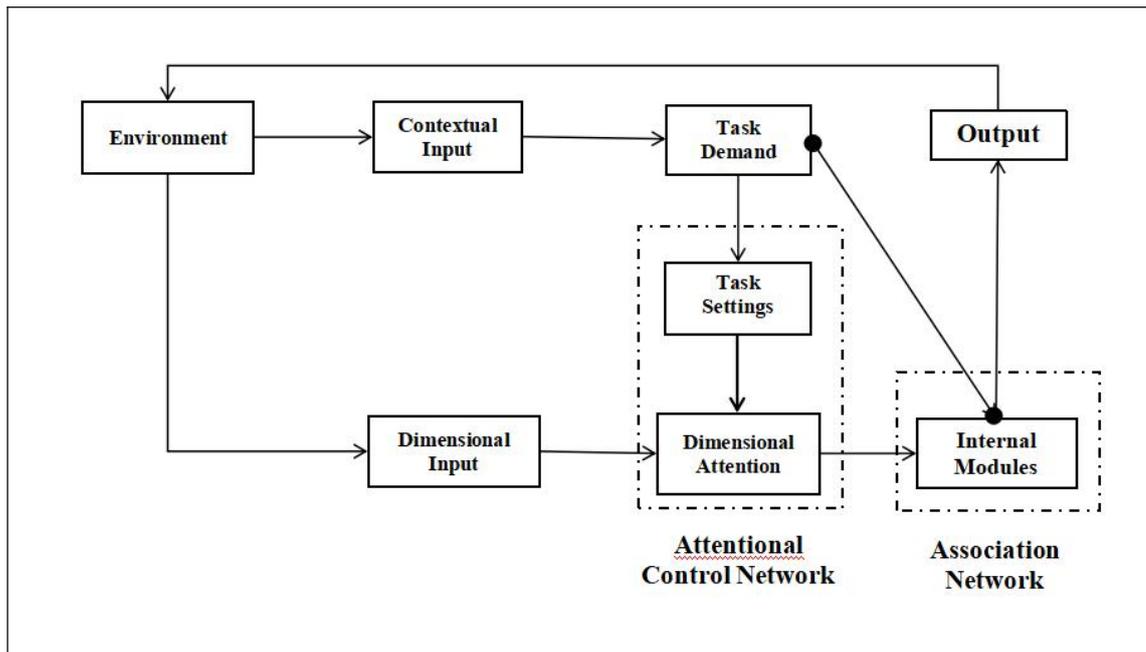


Fig 7-6. A schematic illustration of the integrated modelling framework. The attentional control network reflects the prefrontal control component in categorisation, whereas the association network reflects the striatal control component in categorisation.

More importantly, in this model, it is held that the contextual input and stimulus' dimensional input may have different meanings for human participants. Specifically, contextual input may be more easily associated with the top-down control process, whereas the stimulus' dimensional input is mapped onto the bottom-up influence. This idea is borrowed from studies of contextual modulation of attention in category learning (e.g., Erickson, 2008; George & Kruschke, 2012). In these studies, it was found that extra context cues can modulate the process of selection between internal representation modules.

The next chapter instantiates this theoretical account in a computational model.

Chapter 8.

CATHER: A Combined Model of Task Switching Effects in Categorisation

8.1 Introduction

There is a problem in the traditional computational theory of multiple representations in categorisation. The problem is that most of the models only consider the passive (bottom-up) associative learning process. The bottom-up process even determines higher-level representational attention in the mixture-of-experts network. Although it is plausible that there is no need to take top-down control processes into account within a single category learning paradigm, because this task is merely based on trial and error, in the heterogeneous categorisation paradigm, in which shifts between different sub-representations are essential, the role of attentional control is qualitatively different. In heterogeneous categorisation the costs caused by switching between different sub-tasks are robust. Obviously, the traditional attentional learning account is not sufficient for explaining trial-by-trial task switching (Erickson, 2008). This Chapter introduces a new network model, CATHER (i.e., Control of ATtention to HEterogeneous Representations), that allows the incorporation of task switching theory and multiple representations theory into an integrated modular architecture. The model builds on the theory presented in Chapter 7, where a dual control components account for categorisation was described. The striatal control component corresponds to the association network, whereas attentional control is associated with the PFC. The purpose of establishing the new network

model is, therefore, to computationally instantiate this cognitive control account of categorisation.

8.2 Erickson's (2008) Task Inhibition Proposal

As mentioned in previous chapters, the only model which can implement something like a task switching mechanism is the mixture-of-experts model, ATRIUM. ATRIUM implements its selection of appropriate sub-task through the gating network. In the task switching context, in addition to learned dimensional attention, a higher-level mechanism that modulates selection between subtask modules — internal representations based on different patterns of dimensional attention distribution (e.g., the rule module reflects the representation in which a single dimension is relevant) — is indispensable.

In the original design, the gating network learns to produce a tendency to choose the alternative representations. However, this choice tendency is determined by the learned associations between the exemplar nodes and the gate node, which only models bottom-up processing. It does not include any top-down control influence. In particular, adjustment in the exemplar-based association network as well as the gating network is mediated by dimensional attention, meaning that low-level, dimensional attention drives higher-level, representational attention (e.g., Sewell & Lewandowsky, 2012). This dimensional attention driven control account is problematic. On the one hand, it is inconsistent with the attentional control theory. Although the attentional control theory also includes item-specific influences on control (e.g., Blais et al., 2007; Gilbert & Shallice, 2002), these bottom-up influences are fairly limited. For example, in the Gilbert-Shallice model, the weights between input unit and task demand units are adjusted at the end of each trial. Thus, the weights derived from the activation of the units at the end of trial N only affect the model behaviour for trial $N + 1$. Therefore, according to the attentional control theory, dimensional attention in each module would not affect top-down processing. On the other hand, the dimensional attention driven account may lead to a bias to the use of exemplar-based representations. Consider the original design of ATRIUM, where the rule

module represents attention on a single dimension, while the exemplar module represents attention on multi-dimensions. As ATRIUM holds multiple representations are processed in parallel, with gradient descent on error, it may predict that behaviour would be gradually determined by the exemplar-based representation (i.e., in the original design of ATRIUM, the model indeed implements the dominance of the exemplar module at the end of training; see Chapter 6).

Erickson (2008) acknowledged that, though ATRIUM holds that categorisation can be construed as involving task switching, it does not currently predict switch costs, nor does it have any mechanism that has been identified as representing attentional control. In the traditional categorisation paradigm, the main dependent measure is accuracy. In heterogeneous categorisation, however, RTs become more important. To account for RTs would require the inclusion of principles in the model to govern its dynamics. To develop a new computational account, Erickson (2008) suggested three essential properties to be considered:

- 1). Activation builds up over time;
- 2). It takes time for activations (of task demand units) to decay back to a resting state in the absence of input;
- 3). Activation for one potential response must exceed that of the other responses by some amount for the response to be generated.

A model involving these properties should be able to generate responses more quickly on the second of two consecutive trials that use the same task representation, because on the second trial, residual activation in the control mechanism from the first would allow the selected representation to influence the output more quickly, and this will result in faster responses. Likewise, on a trial that requires a switch from one task to another, it would take time for the control mechanism to allow activation from the other task representation to begin to influence the output, and this would delay the model's response.

In addition to the new properties of the dynamics, Erickson (2008) also recommended two properties of the gating network of ATRIUM to be retained. The first property is the lateral

inhibition between task demand units, and the second is the bias in achieving the multi-dimensional categorisation task. The activations in the gating network are expressed as:

$$a_r^g = \frac{\exp(\phi^g net_r^g + \beta_r)}{\sum_r \exp(\phi^g net_r^g + \beta_r)} \quad (9.1)$$

where net_r^g is the weighted sum of activations in the exemplar network, representing the bottom-up input, $\phi^g > 0$, is the gain of the gating mechanism, which determines how extreme the activation tends to be, and $\beta_r \leq 0$ is the bias to use representation r . For heterogeneous category representations, the gain may have a greater impact on the shifts between representations than for a homogeneous category representation, because it effectively implements the lateral inhibition between multiple representational influences. In contrast, the bias may interfere with multitasking processing, because the bias can push the gating mechanism so as not to (easily) use some alternatives. The bias parameter is necessary to model switching between tasks unequal in difficulty.

Erickson's (2008) proposal is valuable, but he has yet to provide a computational implementation of this account. However, Chapter 7 has already introduced a computational account of task switching based on the Supervisory System theory — the Gilbert-Shallice (2002) model — and argued for its potential for merging into the account of multiple representation categorisation. Interestingly, that model's modular architecture and its residual activation principle, to some extent, are consistent with that proposed by Erickson (2008). However, as mentioned earlier, the gating network mechanism of the ATRIUM model may not be easily mapped onto task switching control, since the exemplar-based representation may not reflect the intrinsic characteristics of task switching control. Instead, the modular architecture of the Gilbert-Shallice model has been verified to be successful in accounting for task switching based on Stroop stimuli, which can be considered as having a heterogeneous category structure (see also Gilbert & Shallice, 2002). Therefore, this Chapter presents an implementation of a new network model that incorporates both the modular architecture of the Gilbert-Shallice task switching model and the multiple representations account, to simulate task switching effects. This computational account may provide a starting point for further investigating trial-by-trial task switching as in the recent literature on heterogeneous categorisation.

8.3 Model Formalisation

8.3.1 Model Architecture

CATHER (stands for Control of ATtention to HEterogeneous Representations) holds three basic assumptions. First, category membership is determined by the similarity between the input stimulus and stored category exemplars in memory. The sub-task modules that simply implement input-output associations are, thus, replaced by the ALCOVE-like feed-forward exemplar-based networks. The association weights between each exemplar nodes and category nodes are established through learning. Second, category learning requires processing multiple representations in parallel which compete to dominate the generation of responses. This idea is borrowed from the multiple representations theory of category learning (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998). This account is also consistent with the attentional control theory. As mentioned in Chapter 7, the original Gilbert-Shallice model also suggested competition between parallel processing modules. Third, and different from some traditional theoretical accounts of the control of competition between multiple representations, the model assumes that the task representation involved in categorisation requires the supervisory system to mediate switching/competition between distinct subsets of internal representations. Thus, in CATHER, on one hand, different internal representations inhibit one another, and on the other hand, the top-down control processes can modulate the inhibition and excitation of internal representations, in particular, in task switching.

The model should be able to perform categorisation tasks with either homogeneous or heterogeneous representation. But, note that, in this chapter, the version of the model is presented in the context of performing the heterogeneous categorisation that needs trial-by-trial switching between a 1D ERB task and an II task.

An illustrative representation of the model architecture is shown in Fig 8-1. As can be seen from the figure, the task demand units and their multiple inputs structure is implemented in the same way as in the Gilbert-Shallice (2002) task switching model. For simplicity, the model

consists of two sub-task modules (i.e., 1D representation module and 2D representation module), and a task switching mechanism with two task demand units. As mentioned earlier, the task demand units reflect the activity in the PFC during task switching. The model thus instantiates the assumption that task switching in heterogeneous categorisation is mediated by a PFC-based Supervisory System.

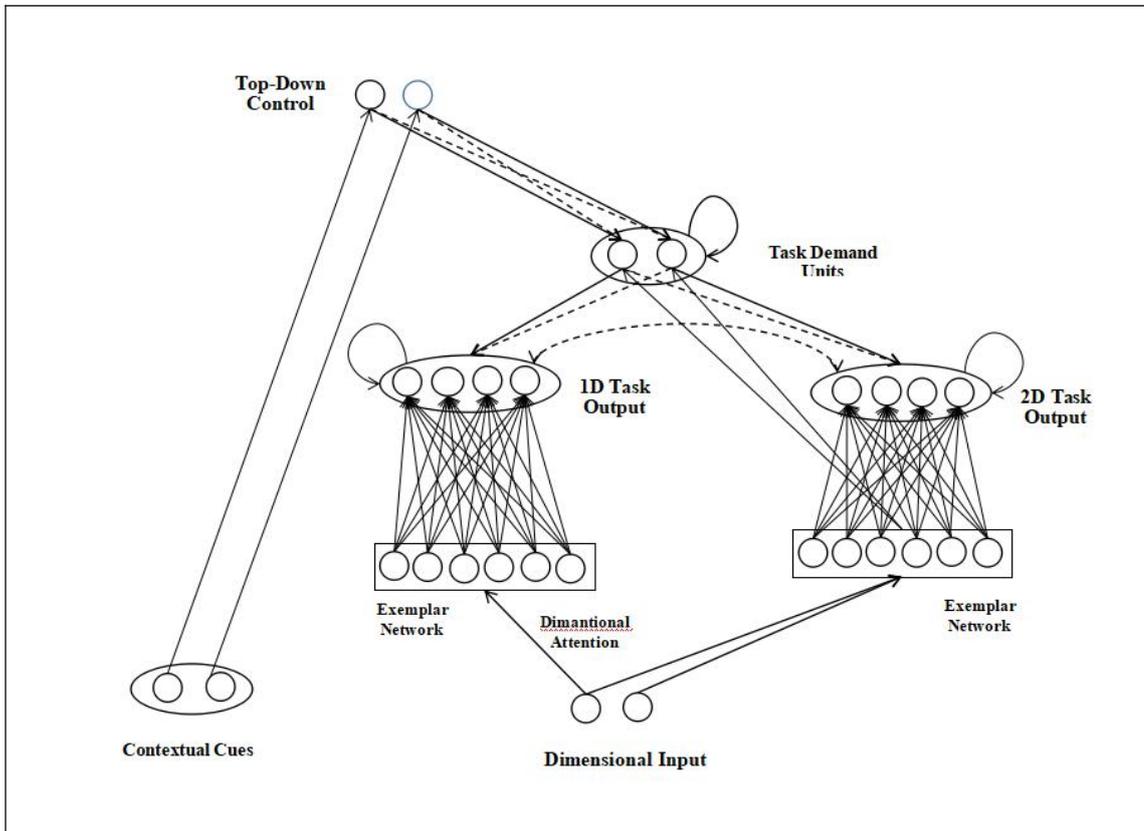


Fig 8-1. An illustrative representation of the network model of task switching in categorisation. The solid bold lines represent excitatory connections, whereas the dashed bold lines represent inhibitory connections. Others lines represent the default or learned association weights.

However, the input representation and the representation of sub-task modules are different from the original task switching network model. First, instead of implementing sets of direct connections between input and output units, the new network model incorporates exemplar-based association networks as representation of the sub-task modules. In each trial, activation levels within the exemplar networks are driven by dimensional attention. The distribution of dimensional attention for different tasks is different. To be more specific, for an ERB task, just one dimension is relevant, thus, the attention strength on one dimension is set to

1.0 and 0 for the other; whereas in the II task, both dimensions are relevant, thus the attention strengths on both dimensions are set to 1.0.

Following the experimental procedure, the model is trained before the task switching phase (though learning may remain after training, the effects become fairly negligible). Association weights in each module are adjusted by the error reduction mechanism. Therefore, instead of the default settings, the strength of the association weights in the modular network represents the strength of remembering of each exemplar-response mapping that has been learned.

Second, the model retains the modifiable connections from input to task demand units, which are used to mimic the influence of item-specific priming, but the input units and their activation levels are represented as an exemplar-based network (Erickson & Kruschke, 1998; 2002). For stimuli used in the recent experiments, which involve continuous stimulus dimensions, an item's receptive field for a given dimension is centred at the presented item's position along that dimension. This builds up a multi-dimensional psychological space. Each item is thus represented as a point in the psychological space. In the present version, this item-specific priming is done by using the exemplar-based network of the 2D representation module.

Others aspects of the model, including lateral inhibition between/within modules, lateral inhibition between task demand units, and excitatory and inhibitory connections between output units and task demand units, are similar those of Gilbert and Shallice (2002).

8.3.2 Model Description

8.3.2.1 Internal Representation Modules

The implementation of each modular association network within CATHER is the same as the original exemplar-based network, except that instead of making the model learn dimensional attention strengths, the dimensional attention strength is set by default. This is to mimic the distribution of dimensional attention in each internal representation module. Moreover, in the task switching experiment, the attentional focus to each dimension was told to the participants before learning each task. For simplicity, two modules were provided within the model to

represent the 1D rule-based task and an II task, respectively. Thus, activation of each exemplar node, a_j^{ex} , is expressed

$$a_j^{ex} = \exp(-c \sum_i \lambda_i d_{i,j}) \quad (8.2)$$

where c is the constant that determines the specificity on exemplar activation, λ_i is the given attention strength on dimension i , and $d_{i,j}$ represents the distance between input stimulus and exemplar j on dimension i . Activation of output unit, a_k^{out} , of each module is, thus, expressed

$$a_k^{out} = \sum_j w_{j,k}^{ex} a_j^{ex} \quad (8.3)$$

where $w_{j,k}^{ex}$ represents the association weights between each exemplar node and category responses.

Before the task switching phase, both modules went through two sessions of training. The association weights in these modules were adjusted by error reduction. In this version of the model, however, following Gilbert and Shallice (2002), learning effects in these internal representation modules become negligible during the task switching phase of the experiment.

8.3.2.2 Model Operation

Processing in the task switching component of CATHER is as follows. At the beginning of each simulation, all units are initialised with zero activation. The top-down input is added to a task demand unit, depending on which task is required. If there is no preparation interval, input units on each dimension are activated. These activations then iteratively propagate throughout the model. On subsequent trials, TD units are set to a proportion (80%) of their activation levels at the end of the previous trial (as in the original task switching model of Gilbert & Shallice, 2002). This models the effects of residual activation. All other units are initialised as zero activation. If there is a preparation interval, all stimulus input units and output units are set to zero, but only top-down input is applied. After the preparation interval, the task demand units are activated as before, and stimulus input units and output units are activated.

8.3.2.3 Task Demand Units and Output Units

Both task demand units and output units receive multiple inputs. In addition to the top-down input that indicates which task it to be performed, the task demand units also receive input from exemplar-based input units, output units, and from the other task demand unit. The net input for task demand unit m in each cycle is expressed as

$$net_m^{TD}(t+1) = S_m^{tdc} + \sum_j w_{j,m}^{ex}(t)a_j^{ex}(t) + \beta_m \quad (8.4)$$

where S_m^{tdc} represents the top-down input for task m , a_j^{ex} and $w_{j,m}^{ex}$ are activation and weights of the bottom-up input units in the II task representational module, and β_m is the bias parameter for each task demand unit. The activation of bottom-up input units follows the exemplar-based representation term. But, unlike the implementation of ATRIUM's gating network, exemplar-based inputs are here used to simulate the short-term item-specific priming effects, i.e., that the effects of learning on trial N persist only for trial $N+1$.

The output units of each module receive inputs from the exemplar network, task demand units, and output units of the alternative module. The inputs from the alternative module reflect that the output units suppress their incongruent alternatives via lateral inhibition. The net input for output unit k in each cycle is expressed as

$$net_k^{out}(t+1) = \sum_n w_{j,k}^{ex}(t)a_j^{ex}(t) + \sum_m w_{m,k}^{TD}a_m^{TD}(t) + \sum_{k'} w_{k',k}^{out}a_{k'}^{out}(t) \quad (8.5)$$

where a_n^{ex} and $w_{n,k}^{ex}$ represent the activations and weights in the exemplar network, a_m^{TD} and $w_{m,k}^{TD}$ are the activations and weights of task demand units, and $a_{k'}^{out}$ and $w_{k',k}^{out}$ are the output units' activations and weights in the alternative module.

8.3.2.4 Activation Calculation

Each trial is ended when the activation level of the most active output unit of one module exceeds that of the most active output unit in the other module by a response threshold. The number of cycles taken for this to occur is then converted to the simulated RTs. The equation for

calculating unit activation in the network model is unchanged from the Gilbert-Shallice (2002) model. The change in activation value for each unit on each cycle, Δa , is calculated as

$$\Delta a = \begin{cases} \partial \times net_i(t) \times (a_{\max} - a(t-1)) + \delta & (\text{if } net_i \geq 0) \\ \partial \times net_i(t) \times (a(t-1) - a_{\min}) + \delta & (\text{if } net_i < 0) \end{cases} \quad (8.6)$$

where a_i is the unit's current activation level, net_i is its net input, δ is a noise term, drawn from a Gaussian distribution, with a mean of 0 and standard deviation of 0.006, and ∂ , a_{\max} and a_{\min} are parameters affecting step size, and maximum (1.0) and minimum (0.0) unit activation values respectively. In the original Gilbert-Shallice model, the minimum unit activation value was set to -1.0, whereas in ATRIUM, the activation of output units and gating unit are in the range between 0.0 and 1.0. Here, I choose the [0, 1] interval to limit the activation level of each unit. In the original Gilbert-Shallice model, a problem of the [-1.0, 1.0] interval is that, sometimes, difference between the most active response units of different modules reached the threshold, but the activation levels may be negative. Thus, this modification converges the range of variation of each unit in the process to avoid some extreme values.

8.3.3 Model Behaviour

8.3.3.1 A Simplified Erickson's Task

For illustrating the model behaviour, in this section, a simplified Erickson's task is introduced. The category structures were the same as those used in Erickson's (2008) (see Fig 6-9). The model training phase was the same as the original experiment. The model was trained for 750 training trials before transfer phase. Thus, association weights within each internal representation module were trained (see Fig 8-2). The trained model was used to process different types of trials (see Fig 8-3). After that, 16 intermixed trials were introduced, the trained model was used to process the intermixed trials (Fig 8-4).

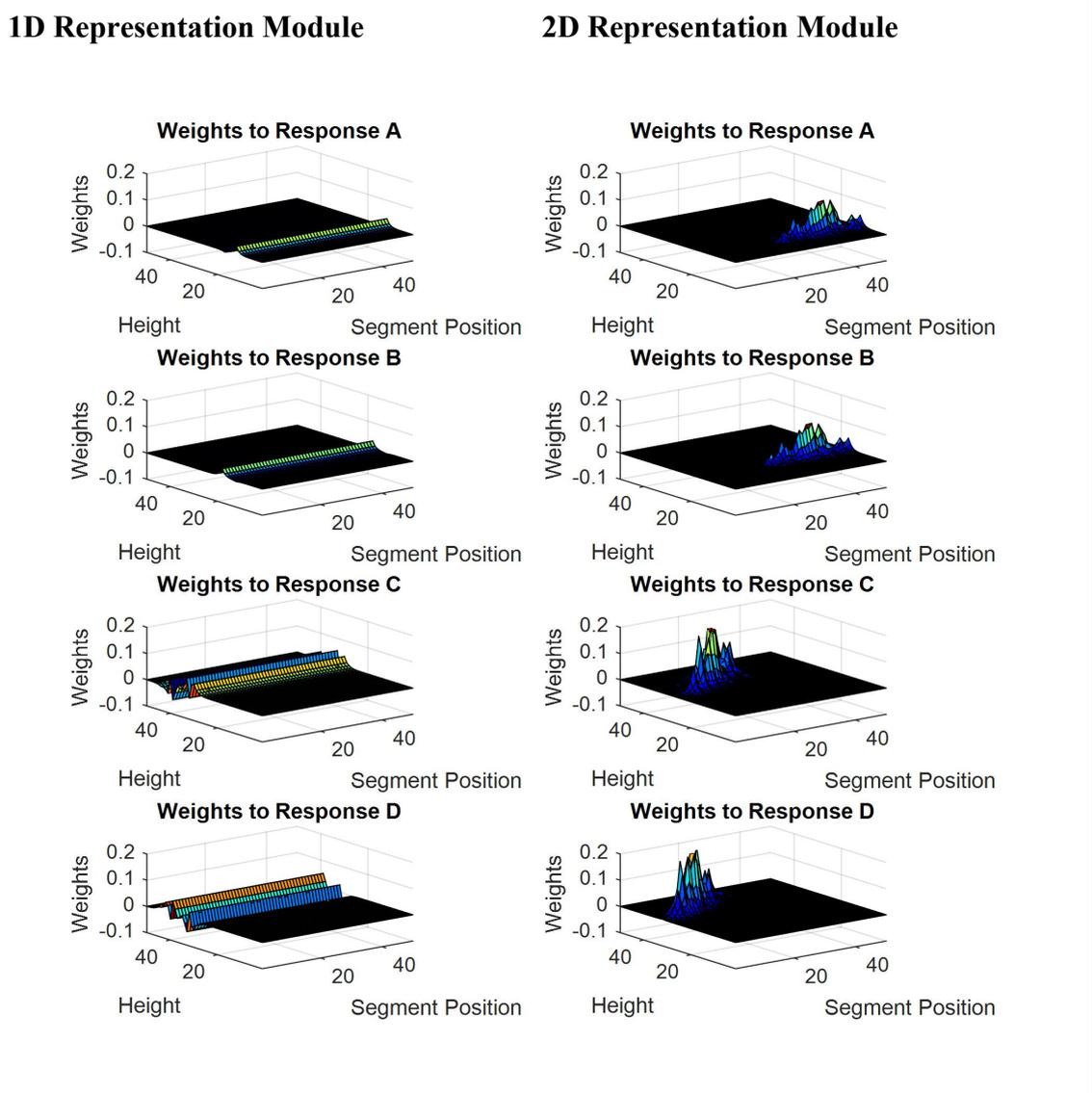


Fig 8-2. The association weights learned throughout the training phase corresponding to Erickson (2008) design. Left panel represents the weights in the 1D representational module, and Right panel represents the weights in the multidimensional representational module.

8.3.3.2 Implementation

Following the experimental design (e.g., Crossley et al., 2017; Erickson, 2008; Helie, 2017), before implementation of task switching, CATHER is trained for hundreds of trials to build up the association strengths between the exemplar nodes and category response units in each module. For simplicity, current version of the model only consists of two representation modules, a one-dimensional (1D) module and a multidimensional (2D) module. Fig 8-2 shows the learned association weights in each module corresponding to Erickson’s (2008) experimental design. In the training phase, the model is trained in 750 trials. As can be seen in the figure, the exemplars

in the 1D module share the same association weights, whereas association weights in 2D module are item-specific.

By using the trainable association network of the internal representation module, we can use trained association weights instead of the weights created by hand-setting. Based on these trained internal representation networks, the CATHER model can now launch the task switching phase. In the task switching context, there are four trial types, two types of stay trials in which the stimulus on the current trial and the stimulus on the preceding trial are from the same task type (i.e., II-II and 1D-1D), and two switch trials in which the stimulus on the current trial and the stimulus on the preceding trial are from different task types (i.e., II-1D and 1D-II). Fig 8-3 shows the example activation profile in each trial type. Prior evidence has shown that the mean reaction time on each trial of the II task is greater than that of the 1D task (e.g., Ashby et al., 2003; Helie et al., 2010; Maddox et al., 2004b; 2004c; Soto et al., 2013), and learning the optimal strategy for an II task normally took a longer period of time than 1D tasks. Thus, it seems that the optimal strategy of the II task is harder to capture than 1D tasks. CATHER is subject to this basic principle. In the task switching phase, the task demand unit of the II task has a greater bias than task demand unit of the 1D task. This can lead to 1) slightly more cycles in two consecutive II trials (173 cycles) than in two consecutive 1D trials (140 cycles), and 2) slightly more cycles for switching from 1D to II task (169 cycles) than from II to 1D task (141 cycles).

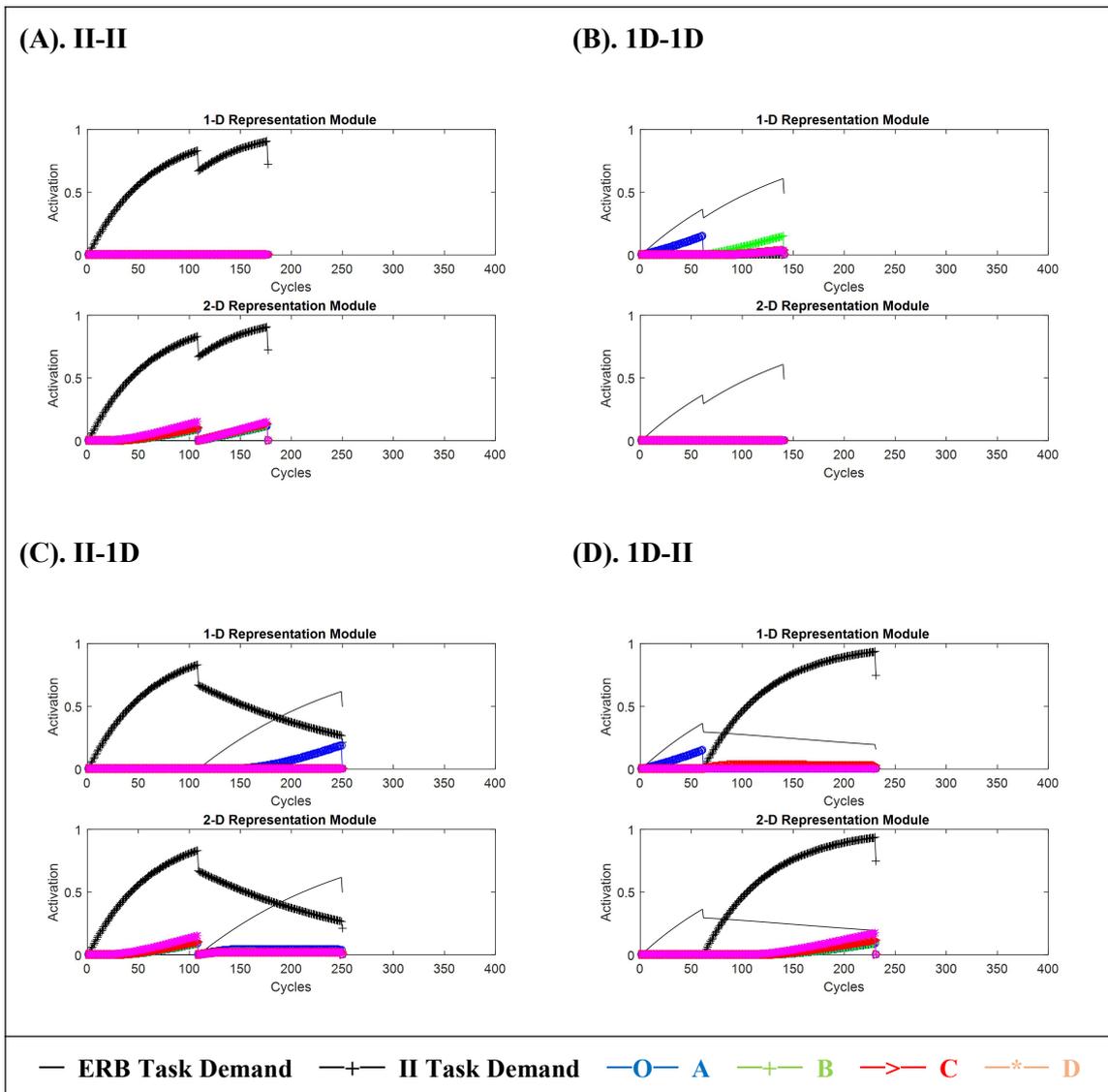


Fig 8-3. Activation profiles of different switch and stay trial types. According to the experimental design, there are four types of trials: two types of stay trial, change from II task to II task (II-II) and change from RB task to RB task (1D-1D), and two types of switch trial, change from RB task to II task (1D-II) and change from II task to RB task (II-1D).

8.3.3.3 Results

Consider a block of intermixed trials. Fig 8-4A shows the activation profiles of each task demand and output units in each module in an example block of intermixed trials. Fig 8-4B summarises the switch effects represented by the number of cycles needed for different trial types. As can be seen in Fig 8-4, not surprisingly, CATHER successfully predicts the switch effects on an intermixed block. The switch trials need more cycles than stay trials (e.g., 161 vs. 80.7).

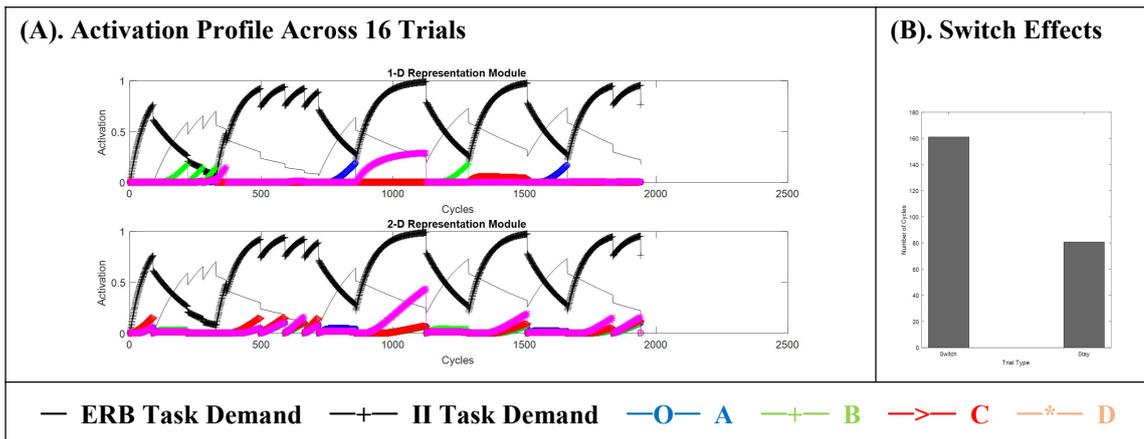


Fig 8-4. Performance of the model in an intermixed block task switching. (A). an example illustration of activation profiles over 16 trials. (B). number of cycles (RTs) averaged across multiple switch and stay trials.

Note that, the current version of CATHER is designed for task switching effects on RTs, but not for accuracy. However, a model considering response errors could be one possibility for extending the model, for which more data sets are needed. The second possibility is to incorporate the threshold in the response decision process.

8.4 Model Applications

The previous section provided the description of CATHER model and illustrated how it is implemented in task switching. In this section CATHER is applied to recent task switching effects in categorisation.

Recent investigations of categorisation have made a connection between switching between multiple representations and task switching (Kiessel et al., 2010). In task switching, participants are typically asked to perform one of two tasks cued on a trial-by-trial basis. Trials where participants need to switch typically suffer from a switch cost in reaction times when compared to consecutive trials using the same task. Switching between categorisation tasks that involve heterogeneous representations would be a special case of task switching in which each task requires different patterns or distributions of dimensional attention (Crossley et al., 2017). Here, CATHER is applied to account for two existing findings: Erickson's (2008) RT costs and Helie's (2017) preparation time facilitation.

8.4.1 Application I: Erickson's (2008) Task Switching Effects

Erickson (2008) conducted the first empirical investigation of control of switching between multiple representations in categorisation. In the experiment, participants were asked to categorise stimuli (rectangles varying in height and position of an internal line segment) into one of four categories (see Fig 8-5A). Two of the categories were distinguished using attention to a single dimension (task 1) while the other two categories required attention to both dimensions (task 2). The category structures used in Erickson (2008) are shown in Fig 8-2A. Each category was associated with a different response button, and tasks were cued using different background colours. It was found that, though the task was very difficult, a significant proportion (40%) of participants were able to perform the combined task, switching on a trial-by-trial basis between task 1 and task 2 as a function of background colour.

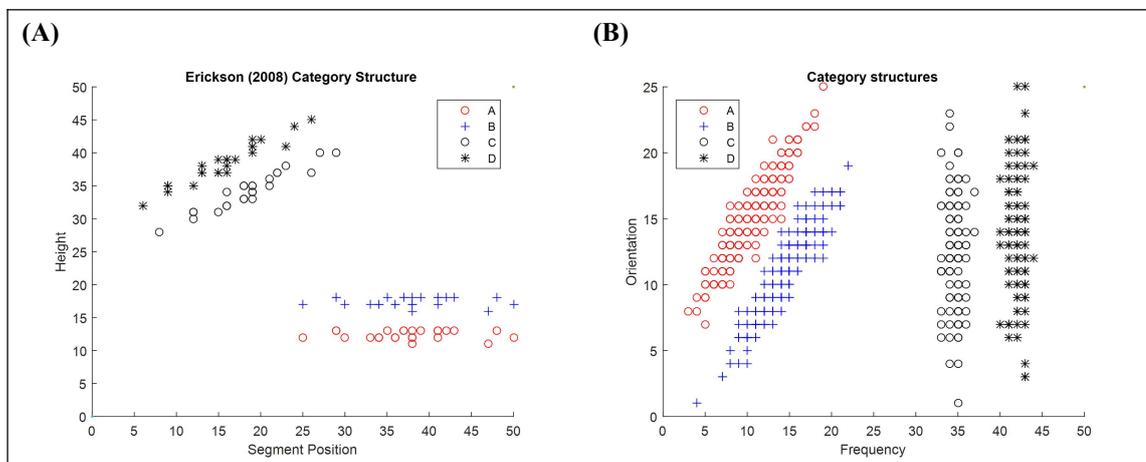


Fig 8-5. Category structures used in (A). Erickson's (2008) and (B). Helie's (2017) Experiments. Stimuli used in Erickson (2008) were rectangles varying on height and segment position, whereas stimuli used in Helie (2017) were Gabor patches varying on bar width (frequency) and bar orientation.

In Erickson's (2008) experiment, participants were first trained on the 1D categories and 2D categories for 200 trials (4 blocks of 50 trials) and 550 trials (11 blocks of 50 trials), respectively. After then, they went through 4 blocks of 100 intermixed trials (i.e., 25 trials for each category). Erickson (2008) proposed that when task partitioning in heterogeneous categorisation occurs,

there should be a cost in response times when participants switch from one sub-task to another. He classified each trial in the last intermixed block as having been preceded by one from the same or the other sub-task. Here, trials preceded by one from the same subtask are referred to as stay trials, whereas trials preceded by one from another subtask are referred to as switch trials. Erickson (2008) found that participants responded significantly more slowly on switch trials ($M \approx 1350$ ms) than on stay trials ($M \approx 975$ ms) (see Fig 8-6, black bars).

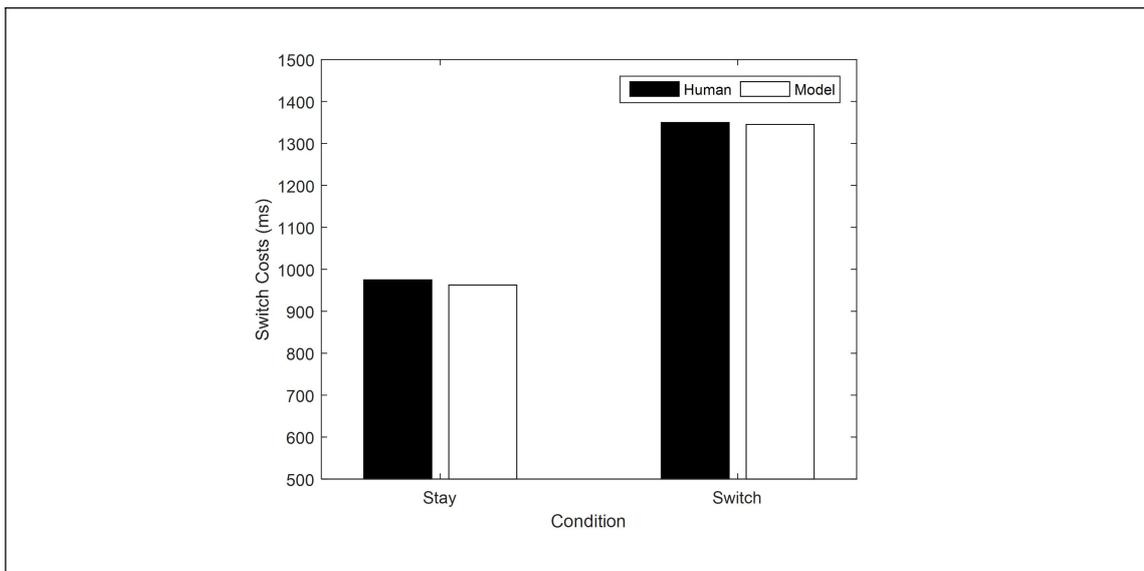


Fig 8-6. Reaction times for stay and switch trials in the task switching phase observed by Erickson (2008) and produced by the model. Simulated RTs = $5.0 \times \text{cycles} + 607$ ms.

8.4.1.1 Method

The CATHER model simulates the task using trial-by-trial presentation of stimuli equivalent to that seen by participants. In the training phase, the modular association networks were trained for 750 trials (i.e., 200 trials for 1D categories and 550 trials for II categories). Note that the purpose of modelling work here is to reproduce RT costs in task switching, and not for fitting training data. (Competition between multiple representations will not be considered in this section, but future directions for extending this network model should be made to take learning processes into account.) After the training phase, the model was run in the task switching phase. As in the experimental design, this consisted of 400 intermixed trials. Only the last 100 trials were used for calculating the RT data, and, following Erickson (2008), stay trials preceded by the same response were excluded from the analysis.

The parameter settings for this simulation are listed in Table 8-1. Two points are noteworthy: 1) it is assumed that multidimensional task is more difficult than the 1D task, so the top-down control input for II task is greater than for 1D task; and 2) following the assumption of the multiple representations theory, the control mechanism should bias the multidimensional representation, and so a bias was added to the net input of the task demand unit of the II task. Others parameters were set similar to the original Gilbert-Shallice task switching model.

Table 8-1
Parameter Settings of CATHER Model

| Parameter | Settings |
|--|----------|
| 1 Exemplar activation specificity | 0.80 |
| 2 Learning rate for weights of 1D representation module | 0.03 |
| 3 Learning rate for weights of 2D representation module | 0.01 |
| 4 Response threshold | 0.15 |
| 5 Step size | 0.0015 |
| 6 Squashing of task demand units | 0.80 |
| 7 Noise | 0.006 |
| 8 Output units bias | -1.0 |
| 9 1D Task demand unit bias | -1.0 |
| 10 II Task demand unit bias | -4.0 |
| 11 Top-down control strength (1D) | 6.0 |
| 12 Top-down control strength (II) | 15.0 |
| 13 Lateral inhibition | 1.0 |
| 14 Between modules inhibitory and excitatory connection strengths | 1.0 |
| 15 Task demand-output inhibitory and excitatory connection strengths | 3.0 |
| 16 Learning rate for adjusting connection weights between stimulus input and task demand units | 1.0 |

8.4.1.2 Results and Discussion

The simulation results are shown in Fig 8-6. Quantitatively, the model did a very good job in fitting the RT data. The model requires more cycles for switch trials (number of cycles = 247.34) than for stay trials (number of cycles = 145.22). For comparing to the empirical data, following Gilbert and Shallice (2002), the simulation cycles are converted to RTs using a linear regression (see Fig 8-3). CATHER is designed to account for RT switch costs, but not to predict errors. Thus, the model may perform much better than human participants on accuracy. But, this

simulation reveals that the CATHER model is able to account for task switching costs on reaction times.

8.4.2 Application II: Helie's (2017) Preparation Effects

Crossley et al. (2017) replicated the task partitioning paradigm of Erickson (2008) to explore 1) if switching between 'explicit' and 'implicit' category learning tasks is possible, and 2) whether switching between 'explicit' and 'implicit' category learning tasks is different from switching between two 'explicit' category learning tasks. They found a significant proportion of participants can successfully perform switching between 'explicit' and 'implicit' categorisation tasks on a trial-by-trial basis. To compare with switching between 1D and II tasks, they used an additional heterogeneous category structure that consisted of a conjunctive 2D rule substructure and a 1D rule substructure, but they found very little difference between these two tasks. This research further confirmed the need for a control mechanism in heterogeneous categorisation.

Helie (2017) explored a factor that has been useful in traditional task-switching paradigm: preparation time. It was found that, when participants familiar with the tasks, their performance benefitted from additional preparation time. According to the decision bound model analysis, after practice, a larger proportion of participants were able to perform switching between RB and II subtasks, and the RT switch costs (i.e., the difference between RT on switch trials and stay trials) were significantly decreased (see Fig 8-7, left bars).

8.4.2.1 Method

Like the Gilbert-Shallice (2002) model, CATHER simulates the effects of preparation time by activating the task demand units in advance of the presentation of stimulus input, without activating output units. The model only simulates the effects of preparation time observed in the second training session of Helie (2017), as the current version of CATHER does not include any learning mechanism in the task switching control network. In this simulation, the preparation interval was set at 50 cycles (equivalent to $3.75 \times 50 = 182.75$ msec in the regression equation). During the preparation interval, only top-down control input was applied to task demand units,

and the task demand units inhibit one another so as to reflect lateral inhibition between units at the same level. Parameter settings are again as shown in Table 8-1.

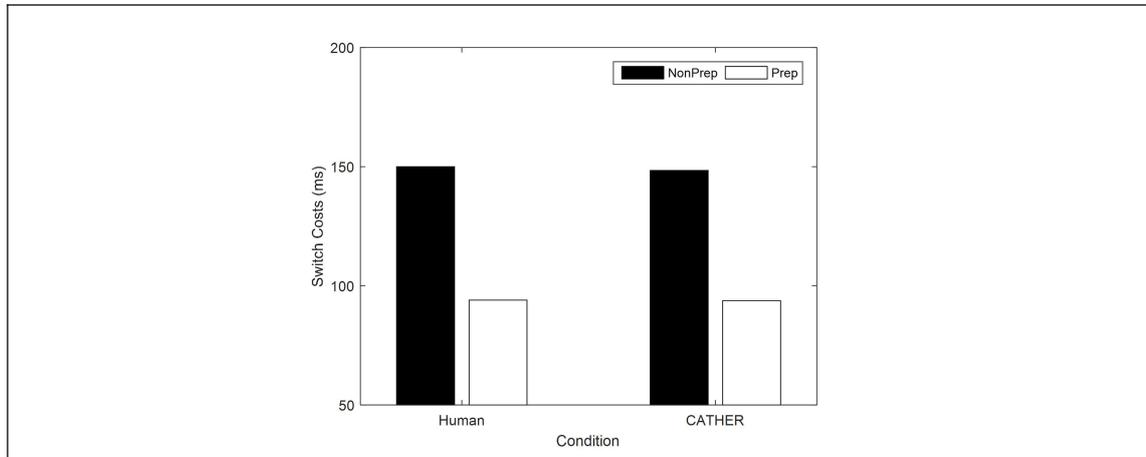


Fig 8-7. Mean RT switch costs in each condition observed in Helie (2017) and predicted by CATHER model. Note switch costs are calculated as the difference between RT on switch trials and stay trials. RT (switch costs) = $1.1 \times \text{cycles}$.

8.4.2.2 Results and Discussion

Results of this simulation are shown in Fig 8-7. As can be seen from the figure, CATHER predicts a great difference in switch costs between non-preparation condition (Mean = 138.02) and preparation condition (Mean = 85.34). Although the regression equation used for calculating RTs differs from that of section 8.4.1, because different empirical studies used different stimuli and training and test procedure, it is sensible to have different patterns of RT costs. Moreover, the focus of this section is on establishing the mechanism which produce the task switch effects observed in the empirical studies, but not on the conversion of results produced by models. In this sense, it is good to see that CATHER can not only predict RT switch costs, but also facilitation effects of preparation time.

8.5 Discussion

8.5.1 Comparison with Earlier Models

The CATHER model combines the modular architecture of a task switching model with the multiple representations theory of category learning, demonstrating the dynamics of attentional control in the representation of heterogeneous categorisation. The dynamics in this model is driven by the interaction between task demand units and modular output units. Activations in task demand units are primarily determined by top-down control inputs, bottom-up inputs (item-specific), and lateral inhibition from the alternative task. The residual activations in task demand units after each single trial are implemented by squashing the activation levels of task demand units to some proportion of their activation levels at the end of the previous trial, as in the model of Gilbert and Shallice (2002). The inhibitory and excitatory inputs from the task demand units strongly interfere with the activations in the modular output units. The residual activations in task demand units can, thus, facilitate the consecutive trials and restrain the switch trials. Not surprisingly, the model successfully predicts the RT costs for different trial types of the Erickson (2008) data set and Helie's (2017) facilitation effects of preparation time.

To my knowledge, CATHER is the first model that produces responses instead of choice probability. Most traditional models, like ATRIUM and SUSTAIN, do not produce a response, but they produce a response probability that is converted, with some non-determinism, to compare with human performance. This is because they were designed for fitting accuracy data sets, but not reaction times. COVIS was designed for response production, but it assumed that cognitive control is not involved in the interaction between different internal representations. Moreover, the only account, EBRW (exemplar-based random walk), that was originally designed for RTs, does not include an associative learning mechanism, and does not account for the role of top-down control.

However, the current version of CATHER is aimed to account for switch effects on RTs but not response errors. CATHER produces responses, and hence the production of error responses needs to be treated differently. One possibility is that top-down unit weights are weakened from those given in Table 8-1. However, this may give rise to a fundamentally different account of

error on categorisation tasks – it makes it part of the output stage rather than part of the categorisation stage.

8.5.2 Parameter Settings and Implications

The parameter settings of CATHER for testing and the applications are shown in Table 8-1. As shown in the table, learning rates for separate internal representation module are different. This idea is borrowed from the traditional modular networks of category learning. Erickson and Kruschke (1998), for example, applied different learning rates for the rule-based module and the exemplar-based representation module. More importantly, following Erickson (2008), the model has involved 1) activation over time, 2) residual activation (i.e., squashing activation), and 3) a response threshold. Moreover, the CATHER model also borrows the idea from multiple representations theory that assumes that tasks based on 1D representation module is much easier than those based on 2D representation module. This is done by adding a greater bias to 2D task demand unit net input than 1D task demand unit net input. In contrast, since in the experimental procedure, human participants were always received longer training on II task than on 1D task, and, according to Gilbert and Shallice (2002), greater effort is needed for the more difficult task, the strength of top-down control input of II task is greater than that of the 1D task. These results imply that, nevertheless, the parameter settings of the current version of CATHER generally reflect the principles of multiple representations theory and task switching.

8.5.3 Future Extensions of CATHER

The purpose of current work is to present a model that can partly reflect the control of task switching in categorisation. Almost all recent investigations have not reported a strong relevance of accuracy costs in the task switching performance, but the RT switch costs are robust. Therefore, the combined network model specifically focuses on the simulation of RT costs, rather than accuracy costs. Thus, one limitation of the current version of CATHER relates to its capacity in response generation. The model's performance may be too good to compare with

human data in accuracy, though, as mentioned earlier, errors might arise when the strengths of top-down control input is weakened. Moreover, this model is not applied to account for processing in the training phase. One possibility to extending CATHER may be to establish a learning mechanism based on the current architecture, and in particular, the learning of top-down control input to task demand units associations and the learning of the selective inhibition of the task demand units. Moreover, CATHER produces responses, and hence the production of error responses needs to be treated differently. This may give rise to a fundamentally different account of error on categorisation tasks – it makes it part of the output stage rather than part of the categorisation stage.

Chapter 9.

Reflection and Conclusion

9.1 Summary

The present research consists of three parts. In Chapter 2, a considerable amount of recent empirical and computational modelling work was reviewed. There has been increasing evidence implying that cognitive control plays a crucial role in category learning. It is particularly noteworthy that the recent research on heterogeneous category learning has provided a preliminary data set for building a new generation of computational models. But before the establishment of a new model, it is necessary to revisit some of the recent influential computational accounts.

The case studies in Chapters 4, 5 and 6 give valuable insight into the three recent network models based on the multiple representations assumption of category learning, SUSTAIN, COVIS and ATRIUM. All the three models are more or less borrowed from the general framework of the attention learning network, ALCOVE, which itself is discussed in detail in Chapter 3. The first case study demonstrated that although the SUSTAIN model embodies the different structures of internal representations on different category learning tasks, the model fails to account for the representation of a typical heterogeneous category structure, rule-plus-exception task. The non-modular architecture of SUSTAIN cannot yield a precise mechanism for interpreting the interaction between internal representations of a category structure. The second case study gave insight into the influential multiple systems theory and the modular architecture of the COVIS model. The COVIS model only provides the mechanism for

interpreting the competition between internal representations in homogeneous category learning. The third case study demonstrated the modular architecture of the ATRIUM model. As a model originally designed for the representation of heterogeneous categories, the modular architecture and the ‘representational attention’ mechanism make ATRIUM very close to a model of task switching. However, the learning algorithm of the gating network and the representation of internal representations need to be modified.

In the final part (Chapters 7 and 8), by combining the modular architecture of category representation with a network model of task switching in which an attentional control mechanism is embodied, a new modular network model accounting for switch costs in heterogeneous categorisation was developed. The representation of heterogeneous categories needs complexity of internal representations and a top-down control mechanism. The modular architecture of task switching does not involve the complexity of internal representations, whereas models of categorisation do not have a top-down control mechanism. In this sense, the two frameworks complement each other.

9.2 Theoretical Implications

9.2.1 Homogeneous and Heterogeneous Categorisation

According to Ashby and his colleagues, formation of representations in categorisation is controlled by two internal systems. The information-integration category representations are controlled by the procedural learning system, which is independent of executive attention, and automatically competes with the explicit learning system. In contrast, Erickson and Kruschke (1998) argued that a higher-level representational attention mechanism is involved in mediating the competition between different internal representations. Recent empirical evidence has supported the view that cognitive control plays a very important role in categorisation. The influence from top-down control is not only related to explicit learning, but it is also important for mediating the interaction between distinct internal representations. This seems to be consistent with Erickson and Kruschke’s proposal. In addition, the investigations of

categorisation automaticity, too, seem not to support the multiple systems account. It is revealed that automatic categorisation appears to be controlled by a system that is distinct from the two internal systems (e.g., Ashby & Crossley, 2012; Helie, 2010; Soto et al., 2013).

In their informal model of cognitive control, Norman and Shallice (1986) assume that automatic behaviour and nonautomatic behaviour are controlled by distinct systems. The nonautomatic behaviour is held to be controlled by a proactive and flexible control system. This is the so-called Supervisory System. This Supervisory System account has been verified to be capable of accounting for task switching phenomena, such as Stroop effects (e.g., Cooper & Davelaar, 2010; Gilbert & Shallice, 2002) and the Wisconsin Card Sorting Task (e.g., Sood & Cooper, 2013). Intriguingly, Erickson (2008) proposed that, at the conceptual level, the Supervisory System account is also related to the control mechanism of interaction between the internal representations of categorisation.

However, most of the traditional empirical studies are not aimed at exploring the role of cognitive control in non-automatic categorisation. The focus of those studies is only on establishing the nature of the interaction between internal representations at a low level. This approach potentially misleads the investigation of cognitive control in categorisation. For example, the case of Ashby and Crossley's (2010) hybrid category learning study is problematic. The hybrid category structure used in their experiments does not have typical heterogeneity, but one of the two categories can be perfectly distinguished by using a simple 1D rule. As using a 1D strategy could easily achieve over 90% accuracy, this may discourage participants from using task partitioning and showing task switching effects. Despite the problematic design, the work of Ashby and Crossley (2010) still contributes to the research on heterogeneous categorisation because it reveals that there exist inhibitory influences between distinct representational modules (see also Crossley & Ashby, 2015). This argument does not conflict with the Supervisory System theory. For example, in the Gilbert-Shallice (2002) model, there are some inhibitory connections between different modules at the output level. Ashby and Crossley (2010) also argued that there should be a control mechanism that mediates competition between separate systems, and that this control mechanism may involve the functioning of frontal cortex. But computationally instantiating this account in a model is fairly difficult at the current point in time, because there are only a few studies that have attempted to explore the relationship

between cognitive control and homogeneous categorisation (e.g., Miles et al., 2014; Paul et al., 2011). More empirical work is needed in this field.

In contrast, studies of heterogeneous categorisation have provided an easier approach for combining things. All three recent data sets (Crossley et al., 2017; Erickson, 2008; Helie, 2017) have shown that task partitioning and task switching in heterogeneous categorisation are robust phenomena. A significant proportion of participants in each of these experiments showed task switching between distinct sub-representations, and the factor borrowed from traditional task switching, preparation time, could also facilitate participants' performance. In a sense, all these cases support the idea that switching between distinct representations occurs and it is plausible that this is controlled by a Supervisory System like top-down control mechanism.

Nevertheless, recent evidence has shown that cognitive control is important for categorisation, no matter whether the nature of category representations is homogeneous or heterogeneous. However, more behavioural studies will be needed for establishing a full understanding of the role of cognitive control in categorisation.

9.2.2 Cortico-Basal Ganglia Loops and Categorisation

Given that nonautomatic categorisation is controlled by a single control system, a natural question to ask is what is the brain structure that underlies that control mechanism? The multiple cortico-basal ganglia loops in our brain is possibly one solution.

The known organisation of multiple parallel cortico-basal ganglia loops is important for both category learning and cognitive control. In cognitive control, it is argued that these loops implement a gating mechanism for action selection that facilitates selecting of the most rewarding actions while suppressing less rewarding actions, where rewards are represented as dopaminergic signals (e.g., Collins & Frank, 2013; Doya, 2002; Frank, 2005). In particular, these gating mechanisms support some higher-level processes, such as working memory updating and maintenance via loops connecting prefrontal cortex and basal ganglia (e.g., Frank et al., 2001; O'Reilly & Frank, 2006; Todd et al., 2008). In category learning, researchers of the multiple representations theory have argued that the formation of distinct internal representations

of categories is mediated by multiple parallel cortico-basal ganglia-pallidal-thalamic-cortical loops (e.g., Ashby et al., 1998; Seger & Cincotta, 2002; Sloutsky, 2010). Studies of patients with Parkinson's disease (PD) have shown that the impaired cortico-basal ganglia function cannot only cause difficulty in rule-based tasks (e.g., Bowen et al., 1975; Ell et al., 2006), but also impair performance in information-integration tasks (e.g., Filoteo et al., 2005; Knowlton et al., 1996). Based upon the existing evidence, it is argued that the cortico-basal ganglia loops play an important role in the organisation of multiple internal representations in categorisation.

Within the cortico-basal ganglia loops, it is commonly argued that the frontal cortex, especially the prefrontal cortex, should be the central executive component which implements the top-down control processes, while the basal ganglia play a crucial role not only in reward-driven learning, but also in, sometimes, facilitating prefrontal activity in task switching (e.g., Moustafa et al., 2008) and working memory (Cools et al., 2007; O'Reilly & Frank, 2006). Erickson (2008) suggested that the principles of Norman and Shallice's (1986) Supervisory System theory share some properties with the top-down control mechanism in category learning. As is known, the Supervisory System theory addresses the functional architecture of the central executive component (Shallice et al., 2008). Moreover, for answering the question of how the top-down control component intervenes with category learning, Ashby and Crossley (2010) proposed that the prefrontal control may influence response generation in categorisation via the basal ganglia through the hyperdirect pathway. These two ideas together motivate the computational modelling work in this thesis.

In the research on homogeneous categorisation tasks, such as rule-based and information-integration category learning, considerable evidence of the importance of cortico-basal ganglia loops has already been confirmed. But, how the cortico-basal ganglia loops implement task partitioning in heterogeneous categorisation remains unknown. At least, recent evidence suggests that the idea of multiple systems theory (specifically COVIS) – that cognitive control does not influence information-integration category learning – may be wrong. More neuroimaging studies will be needed to confirm and advance understanding of the neural mechanism of top-down control in the organisation of multiple internal representations of categorisation.

9.2.3 The Modular Architecture and Task Inhibition

The role of top-down control in categorisation has long been underestimated. Ashby and his colleagues argued that cognitive control is not involved in the interaction between the explicit learning system and the procedural learning system (Ashby et al., 1998), whereas Erickson and Kruschke (1998) argued that interaction between internal representation modules is coordinated by the bottom-up, exemplar-based gating network. However, as more and more evidence is provided, it is argued, here, that the importance of top-down control in categorisation should be reconsidered.

In the cognitive control theory, it is traditionally assumed that human behaviour should be driven by a series of low-level, internal representations, and the role of top-down control is to mediate the organisation of these internal representations. Thus, we can inhibit inappropriate representation, and task appropriate representation in response to the task demands. The modular architecture of category learning shares some similarities with the internal representations account.

The multiple representations account has been widely accepted in computational modelling of category learning (e.g., Ashby et al., 1998; Nosofsky et al., 1994b; Erickson and Kruschke, 1998). In particular, in previous case studies, it has been shown that the modular architecture of ATRIUM does show some advantages over the nonmodular networks, such as ALCOVE and SUSTAIN. Neither ALCOVE nor SUSTAIN could account for the formation of heterogeneous category representations. Moreover, compared with the modular architecture of COVIS, ATRIUM, especially the representational attention account, seems more flexible.

CATHER embodies the top-down task inhibition into the modular architecture borrowed from earlier category learning models. The task inhibition mechanism (also known as backward inhibition) (Allport et al., 1994) is the hypothesised form of cognitive control in the multitasking environment, arguing that the task switching effects is due to the need to activate a previously inactive, but currently relevant internal representation, and inhibit a previously active, but currently irrelevant internal representation. Task inhibition has been verified to be important in

the control of multitasking situations, such as n-2 repetition (see Kiesel et al., 2010, for a review). The idea is that, given that perceptual categorisation behaviour consists of multiple, competitive internal representation modules, the inhibitory processes should facilitate the use of relevant internal representations to produce responses in categorisation tasks. In particular, this account should be helpful for explaining the task switching effects observed in the heterogeneous categorisation tasks.

The CATHER model assumes that there exist multiple internal representations (e.g., single dimension-based and multiple dimension-based) in categorisation. The top-down control mechanism learns to select the appropriate internal representation in response to task demand. For a homogeneous categorisation task, the control mechanism gradually learns to select one of the internal representations to dominate response generation. In contrast, for a heterogeneous categorisation task, top-down control learns to select between internal representations that are associated with the appropriate subset of the category representation. The version of CATHER presented in Chapter 8 can account for the task switching effects observed in Erickson (2008) and the effects of preparation time observed by Helie (2017). The success of this task inhibition account supports the idea that top-down control should play an important role in categorisation. An integrated account of category learning should combine the modular architecture of category learning and the task inhibition mechanism.

9.3 Limitations and Future Directions

The purpose of this research is in establishing a connectionist model that combines top-down control and multiple representations of category learning to account for the observed task switching effects in heterogeneous categorisation. Thus, the alternative models included in the case studies are all network models (i.e., ALCOVE, SUSTAIN, COVIS, ATRIUM, and the Gilbert-Shallice model). Others models based on multiple representations of categorisation, such as RULEX (Nosofsky et al., 1994) and rational model (Anderson & Betz, 2001), are not considered. This is by no means to say that those non-connectionist models are not worthy of

consideration in the theoretical integration, and future work could give insight into the non-connectionist approaches.

In addition, the version of CATHER described here implements learning in the internal network modules, but it has not yet implemented learning of the cognitive control system. The model has incorporated some principles proposed by Erickson (2008) and Ashby and Crossley (2010), but there remain some other proposals that could be included in the future. For example, Helie (2017) proposed that the cognitive control system in categorisation can be trained, because in Helie's (2017) experimental design, two sessions of training were included, and he did not observe a significant decrease on switch costs until the second training session. Intuitively, it may be easier to switch between two well-known tasks than switching between two unfamiliar tasks. The additional training is beneficial for improving participants' familiarity of the categorisation tasks. Therefore, it would seem appropriate to incorporate a learning mechanism for the control mechanism into CATHER in the future. Presumably such a mechanism would adjust the strength of top-down connections or task-demand units in response to increasingly strong categorisation in the internal representation modules. The third limitation of the current work is that the CATHER model is designed to produce RT costs, but not to produce response errors. CATHER as described here is too accurate. Future extensions of the model should embody a response production mechanism that can be used to fit the accuracy data.

9.4 Conclusions

Through combining two previous modular network models — category learning and task switching — we have produced a novel model of heterogeneous category representation. Instead of adding an additional control module in the modular network of categorisation (Helie, 2017), the model demonstrates the viability of the proposal that a top-down control process of (sub)task switching in heterogeneous categorisation is triggered by competition between internal representations. In the model, on each trial, a top-down process determines which task is undertaken. When (sub)task switching, the top-down control mechanism inhibits the previously active internal representation module and excites the appropriate module. This takes longer than

when a (sub)task repeats. The model reproduces the switch cost effects observed by Erickson (2008), and the effects of preparation time observed by Helie (2017).

Appendix A.

A Preliminary Behavioural Study

A.1 Background

In COVIS, as mentioned earlier, one of the core assumptions is that II category learning may be independent of attention. The Weather Prediction Task (WPT) is one of the well-known II category learning task (Knowlton et al., 1996). In a typical WPT experiment, participants are required to make predictions of good or bad weather according to the presence or absence of four tarot cards. Each card was associated with each possible weather icon with a fixed probability. The overall probability of each outcome on a given trial is calculated according to the conditional probabilities of each weather outcome and card occurring together. Participants are commonly told that they would begin with guessing but they could improve their performance by using feedback.

In a classic study, Knowlton et al. (1996) compared performance of patients with Parkinson' disease (PD) and aged-matched controls in the WPT. They found that, after the first 50 trials, the controls quickly learned the task and achieved an accuracy of about 70%, while the PD patients failed to learn the task and only achieved about 55% accuracy. Knowlton et al (1996) proposed that the failure of PD patients on the task resulted from their striatal dysfunction. In COVIS, the striatum is supposed to be the core structure that mediates II category learning. Therefore, Knowlton et al.'s (1996) result becomes the fundamental evidence of the neurobiological theory.

As mentioned earlier, one piece of supporting evidence that II category learning is independent of cognitive control comes from the investigations of dual task interference in category learning. The dual task interference paradigm, using a dual working memory task, such as the 2-back task, allows the assessment of learning performance under the absence of working memory. Traditional dual task interference studies of category learning showed that secondary tasks that tax executive functioning interfere more with simple rule-based category learning than II category learning (e.g., Waldron & Ashby, 2001; Zeithamova & Maddox, 2006; 2007), suggesting that executive functions are not important for II category learning. Similarly, secondary tasks during categorization feedback interfere with simple rule-based category learning but not II category learning (Maddox, Ashby, et al., 2004; Zeithamova & Maddox, 2007). In the sense, as an II category learning task, the dual task interference should also have no effect on the learning of WPT.

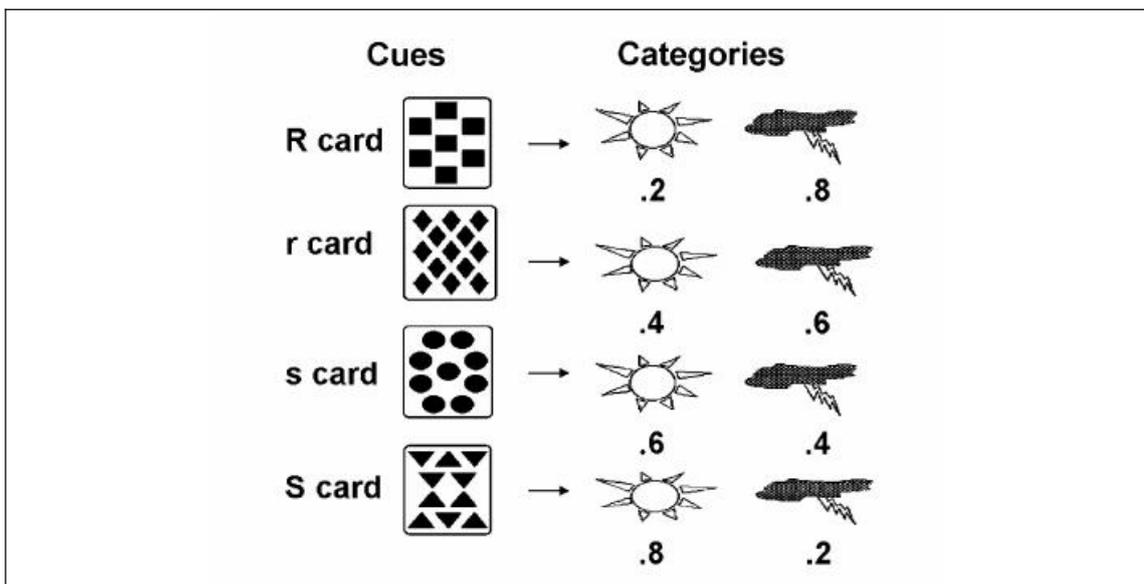


Fig A1. An illustrative example of four cards and the likelihoods with which they predict the outcomes, rain and sun in WPT. The strong rain (R) and sun (S) cards each predict the weather (rain or sun) with 80% probability, while the weaker rain (r) and sun (s) cards each predict the outcome with 60% probability. One, two or three cards are presented on each trial, and the probability of each outcome on a given trial is a function of the probabilities of all cards present on that trial

However, in contrast to the early dual task interference studies, recent research on patients with prefrontal damage, older adults, and children suggests that II category learning may not be fully independent of cognitive control. How can we interpret these discrepancy? There are at least two explanations. First, it may be due to the failure to consider the timing of interference

effects. In traditional dual task interference experiments, cognitive control is disrupted while the categorisation decision was being made or during feedback presentation (e.g., Waldron & Ashby, 2001; Miles & Minda, 2011; Zeithamova & Maddox, 2006). In these cases, the interference was temporary, because executive functions were occupied for a portion of each trial but were available in between trials. Second, it may be due to the failure to consider which elements of executive functioning have been taxed in traditional dual task interference studies. Cognitive control is known as a complex system that is determined by several elementary functions. An influential theory is Miyake et al.'s (2000) three functions theory. In this theory, it is assumed that the executive functioning is composed of an active memory or updating element, a monitoring and response inhibition element, and a set shifting element. Miyake et al. (2000) proposed several executive elementary tasks which are assumed to tap to these three functions. Although some argued that this theory has its weakness in the demonstration of the interaction mechanism, this theory and research of those elementary tasks remains very popular in the domain of the investigation of cognitive control.

In the present study, I introduced three auditory-vocal tasks that fully tax separate elementary executive functions to test 1) if II category learning still occurs when fully taxing executive functions, and 2) if taxing different elementary executive functions can cause differences in interference effects. Three auditory-vocal tasks tapped the three distinct executive functions proposed by Miyake et al. (2000). Cooper et al. (2012) has successfully found differential interference effects on a well-known rule selection categorisation task, the Wisconsin Card Sorting Task (WCST), using this method.

A.2 Method

A.2.1 Participants

Twenty-seven participants were recruited through the department's Sona-system participant panel. To ensure that only participants who performed above chance were included in the data

analysis, a learning criterion of 50% correct during the first 50-trial block was applied. Using this criterion, we excluded 5 participants.

A.2.2 Procedure

Participants all went through four experimental sessions. Session 1 was a control session and sessions 2-4 were dual task interference sessions. In session 1, participants were required to complete 3 50-trial blocks of WPT without time limit or secondary task interference. Gluck et al. (2002) suggested that a problem with Knowlton et al. (1996) is that the association probabilities between the weak cards and weather outcome is too close to chance (57%). Therefore, here I introduced Gluck et al.'s (2002) design in which the four cards were associated with weather outcome with probabilities of .8, .6, .4 and .2, which meant that an individual who always responded with the most likely category for each pattern could correctly predict the weather on up to 83% of trials (compared with 76% under Knowlton et al.'s (1996) design). In the three auditory-vocal interference sessions, participants were required to complete the 3 50-trial blocks while simultaneously completing each of three secondary auditory-vocal tasks. The order of completion of secondary tasks was counter-balanced across participants, with all participants completing each task. In all interference cases, practice on the secondary tasks was given prior to performance of the dual-task conditions. The secondary tasks continued for as many trials as needed for completion of each experimental session. Reaction time (RTs) and responses for each trial were recorded. The manipulation of this experiment was not time limited. So as to avoid machine-related interference between concurrent tasks, one PC was used to administer the WPT and a second was used to administer the auditory-vocal tasks. Participants sat at a comfortable distance in front of the monitor attached to the PC that administered the WPT and interacted with that PC through a mouse controlled by their preferred hand. In sessions 2, 3 and 4 they wore noise-reducing headphones through which auditory stimuli were presented and directed their vocal responses to a microphone positioned in front of the monitor. The experimenter sat beside the participant and manually recorded all responses to each auditory-vocal task.

A.2.3 Secondary Tasks

The three auditory-vocal tasks were the categorisation stop task, which is assumed to tap the response inhibition function, the 2 back task, which is assumed to tap the active memory or updating function, and the digit switching task, which is assumed to tap the set shifting function (e.g., Cooper & March, 2012).

Categorisation Stop Task (CS) Participants heard a series of nouns at a rate of one noun every 2.5 seconds. Their task was to respond vocally with ‘yes’ if the noun was a type of food and ‘no’ otherwise. On one in six trials the noun was followed by a tone, and on these trials, participants were required to withhold their responses (say nothing).

2-Back Task (2-Back) Participants heard a series of digits (in the range 1 to 9, with each digit being equally likely) at a rate of one digit every 2.5 seconds. They were required to respond vocally with “yes” if the current digit was the same as that two trials before. The dependent measures were accuracy (the number of hits and correct rejections divided by the total number of trials) and sensitivity (d' , calculated according to standard principles of signal detection theory).

Digit Switching Task (DS) Participants heard a series of digits (either 1, 2, 3, 4, 6, 7, 8 or 9) at a rate of one digit every 2.5 seconds. Participants were initially required to respond “high” if the digit was greater than 5 and “low” otherwise. After 4 trials of this form, a tone was presented indicating that the required responses had changed and that participants were to respond “odd” if the digit was 1, 3, 7 or 9 and “even” otherwise. Tones were presented after every 4 trials throughout the task to indicate that a switch between the two response sets was required. The dependent variable was accuracy, i.e., the number of correct trials divided by the total number of trials.

A.3 Results and Discussion

A.3.1 Accuracy Based Analysis

To explore the initial learning phase, a 5 x 4 repeated measures ANOVA was conducted with Block (1 through 5 for the first 5 10-trial blocks) and Condition (Control vs. CS vs. 2-Back vs. DS) as within subject variables and proportion of correct responses as the dependent measure.

The ANOVA revealed a significant main-effect of Block [$F(4, 252) = 2.874, p < .05$]. There was also a significant main-effect of Condition [$F(3, 252) = 5.840, p < .01$]. The ANOVA also revealed a significant Block x Condition interaction [$F(12, 252) = 2.363, p < .01$]. In other words, participant accuracy differed across the various conditions (with performance being best on the control condition as shown in Fig A2), and changed across blocks (falling slightly), but the change across blocks was different for the different conditions, with greater improvement appearing to occur under the control condition, but no apparent difference between the three secondary task conditions. To explore the effects of Condition, Block and Condition x Block interaction in the three secondary task conditions, an additional 3 x 5 ANOVA was then conducted. As expected, there was no significant effect of Condition [$F(2, 42) = .310, p = .735$], Block [$F(4, 84) = 1.544, p = .197$], or Condition x Block interaction [$F(8, 168) = .651, p = .734$].

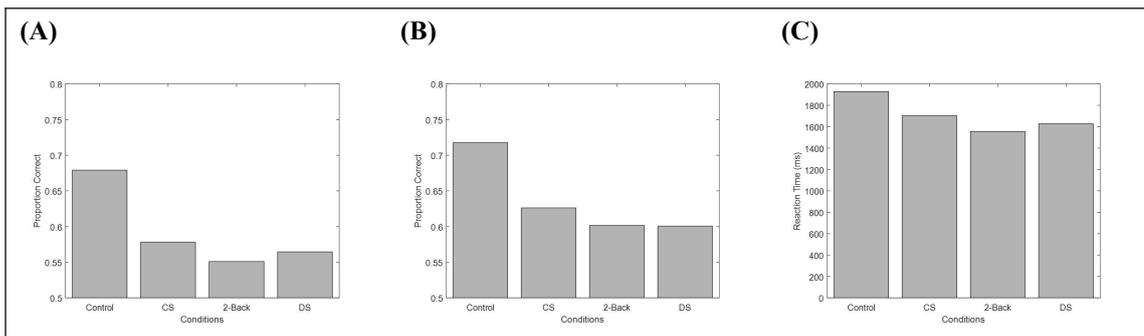


Fig A2. Proportion of correct responses for the first 50 trials (A) and each whole session of 150 trials (B), and (C) mean reaction times (RTs).

A 3 x 4 repeated measures ANOVA, with block (ranging over the 3 50-trial main blocks) and condition as within-subjects variables was then conducted on the accuracy data to explore effects of secondary task on performance in all blocks. The ANOVA revealed a significant main-effect of Block [$F(2, 42) = 18.762, p < .01$], and a significant effect of Condition [$F(3, 63) = 7.477, p < .01$]. However, there was no significant effect of Block x Condition interaction [$F(6, 126) = .275, p = .948$]. In a 3 x 3 ANOVA with Block x Condition (three secondary task interference conditions) as within subject variables (and ignoring the data from the single-task condition), only a significant effect of Block was found [$F(2, 42) = 7.478, p < .001$]. The accuracy-based performance is illustrated in Fig A2.

A.3.2 RTs

A 5 x 4 repeated measures ANOVA was conducted with Block (1 through 5 as the first 5 10-trial blocks) and Condition as the within subject variables and RTs as dependent measures. The ANOVA revealed a significant main-effect of Block [$F(4, 84) = 7.921, p < .01$], whereas there was no significant main-effect of Condition [$F(3, 63) = .901, p = .448$]. There were no other significant effect found in this analysis. Finally, a 3 x 4 ANOVA was conducted with Block (1 through 3 as the 3 50-trial blocks) and Condition as the within subject variables. A significant main-effect of Block was again found [$F(2, 42) = 17.606, p < .01$]. There was no significant effect of Condition or Block \times Condition interaction.

A.3.3 Strategy Based Analysis

One 4 x 3 x 4 repeated measures ANOVA was conducted with Condition, Block (1 through 3) and Cue (with predictiveness of .8, vs. .6, vs. .4 vs. .2) as within subject variables and rate of prediction to category 'Sun' as the dependent measure. There was a significant main-effect of Cue [$F(3, 63) = 58.735, p < .01$], whereas effects of Condition and Block were not significant. There was a significant interaction between Cue and Condition [$F(9, 189) = 7.843, p < .01$], as well as an interaction between Cue and Block [$F(6, 126) = 12.964, p < .01$]. The interaction of Condition x Block x Cue interaction was not significant [$F(18, 378) = .905, p = .573$]. A further 3 x 3 x 4 ANOVA with the three within subjects variables (excluding the control condition) was conducted. A significant effect of Cue [$F(3, 63) = 25.890, p < .01$] and a significant Cue x Block interaction effect were found [$F(6, 126) = 9.921, p < .01$]. There was no other significant effects found here.

A.3.4 Theoretical Implications

The results of this experiment revealed that the three secondary tasks which were assumed to tap to different elementary executive functions had no difference of interference effects either on the pattern of initial learning (the first 5 10-trial blocks), or on the pattern of performance on the 3 50-trial blocks. The learning effect appeared to occur under the interference conditions

across the 3 50-trial blocks, whereas the learning seems not to happen during the early stage of learning under any of the interference conditions. In addition, the analyses based on cue predictions revealed that participants' predictions of the most predictive cue (80%) seems to increase and the least predictive cue (20%) decrease during learning. The interference produced worse performance on the cue predictions. In contrast, the two equivocal cues (40% and 60%) seem to be ignored and the patterns of performance on control and interference of secondary tasks were quite similar.

These results suggested that there is no differential effect of different assumed elementary executive functions. One possible reason may be that the dual-task interference paradigm is too crude. Alternatively, the different functions may affect what is learned, but these differences may not be shown here because accuracy level is a single coarse measure. It is possible that the same accuracy level is produced by the use of different strategies resulting from the different dual-task condition, so in the CS condition people seem to be better at learning that the cue with 20% predictiveness means not 'Sunshine' [$M = .254$, $SE = .157$]. Whereas, it is consistent with the hypothesis that the learning still occurs even under the secondary task conditions.

The dual task interference effects were consistent with the prediction according to Craig and Lewandowsky, (2012) that the learning performance was impaired by each of the three dual tasks conditions. In the meantime, differences have not been found among different interference conditions. In addition, learning still occurs under the dual task interference conditions, but the efficiency or the learning rate was significantly slower than in the control condition. In this sense, it is reasonable to suppose that in a probabilistic category learning task, the deliberative system--cognitive control does play a very critical role, but the specific roles of cognitive control during this processes need to be further explored. Because as was mentioned above, the WPT was widely considered to primarily involve the implicit system, this direct evidence of the involvement of explicit system suggests that the explicit system and implicit system interactively remain during the implicit learning task. On one hand, the explicit system may bootstrap the learning, on the other hand, the procedural reinforcement learning mechanism may help people to slowly, implicitly, or even passively associate the perception to actions.

A probabilistic category learning task or implicit task is characterised by higher uncertainty compared with a rule-based, explicit learning task. In explicit learning tasks primarily requiring

attention and working memory, it has been well established that people can actively monitor the ongoing cognitive status of processing in these systems. However, the understanding of the relationship between conflicting monitoring and the nondeclarative memory system remains very vague. Conflicting monitoring is a very important component of cognitive control (Botvinick et al., 2001). Paul et al., (2011) proposed that explicit learning tasks and implicit learning tasks may share similar a conflict monitoring mechanism. This theory is somewhat consistent with our findings that the probabilistic learning performance under dual task interference were severely impaired. However, this impairments may not be just produced by impaired conflicting monitoring, the implicit learning system may in fact share the whole cognitive control system with explicit learning, but the extent of involvement of cognitive control in these tasks may be different.

Nevertheless, the theoretical implication is that the cognitive control does play some role in probabilistic category learning, although the role is not as important as that in explicit learning. This flexible and active involvement of cognitive control may be the link which combines these systems and dominates the interactions between them.

A.3.5 Limitations

However, it is noteworthy that this experiment has three limitations. First, as was mentioned, the dual task interference paradigm may be too crude. The use of auditory-vocal secondary tasks is a cross modal design which may involve more complexity than the single modal design (e.g., Waldron & Ashby, 2001; Zetthamove & Maddox, 2006). Second, the cross modal design may lead to multiple goals or goal confusion for participants. Third, this experiment was not time limited. According to Ashby and his colleagues (e.g., Ashby et al., 2003; Ell & Ashby, 2006; Maddox et al., 2004a; 2004b; 2005; 2007a; 2007b), each trial within the category learning experiment is self-paced, typically, with an upper time limit of 5 seconds. For example, if a response was not given in that time period, the participants were prompted to speed up their responses, and that trial was discarded. But the self-paced time limit was not included in this experiment.

Appendix B.

Attention Learning Mechanism and Wisconsin Card Sorting Task

B.1 Introduction

Rule shifting is a fundamental prerequisite for categorisation. In terms of the attentional learning theory, a rule refers to a set of dimensions receiving more attention than others. Therefore, the rule shifting is indeed the process of attention allocation. The attention is assigned to predictive dimensions, and away from uninformative dimensions. Kruschke (1992) developed a network model of category learning, assuming that attention allocation in categorisation is acquired by the error-driven learning. The key assumption of the model is that 1) category membership of an item is determined by the similarity of this item to stored exemplars of this category, and 2) learning is achieved by gradient descent on error. This error-driven learning approach has been widely applied in the area of cognitive modelling of category learning. A range of phenomena in research of category learning have been successfully accounted for by attentional learning theory (e.g., Erickson & Kruschke, 1998; Kruschke, 1996a; 1996b; 2001; Kruschke & Johansen, 1999; see Kruschke, 2011 for review). Most of these phenomena are based on the classic category learning paradigm. In a classic category learning task, participants receive a long training session. The training session always consists of hundreds of trials, which allows the gradient descent of error before asymptotic performance. However, when the context requires an individual to make quick and frequent shifts, if the error-driven learning remains plausible?

B.2 The Wisconsin Card Sorting Task

The task we use for testing the model is a simplified Wisconsin Card Sorting Task (WCST). The WCST is a popular clinical and cognitive measure of categorisation ability and executive functioning. In the task, the experimenter has a deck of cards with a variety of figures displayed on each card. The cards differ in the shape, number and colour of the figures (see Fig 1). Each one of these dimensions has four possible values (for a total of $4 \times 4 \times 4 = 64$ possibilities). The simplified WCST consists of the 64 trials. On each trial, the participant is shown a card and asked to sort it using one of the dimensions. After 10 consecutive correct categorisations, the rule (relevant dimension) is changed without instruction.

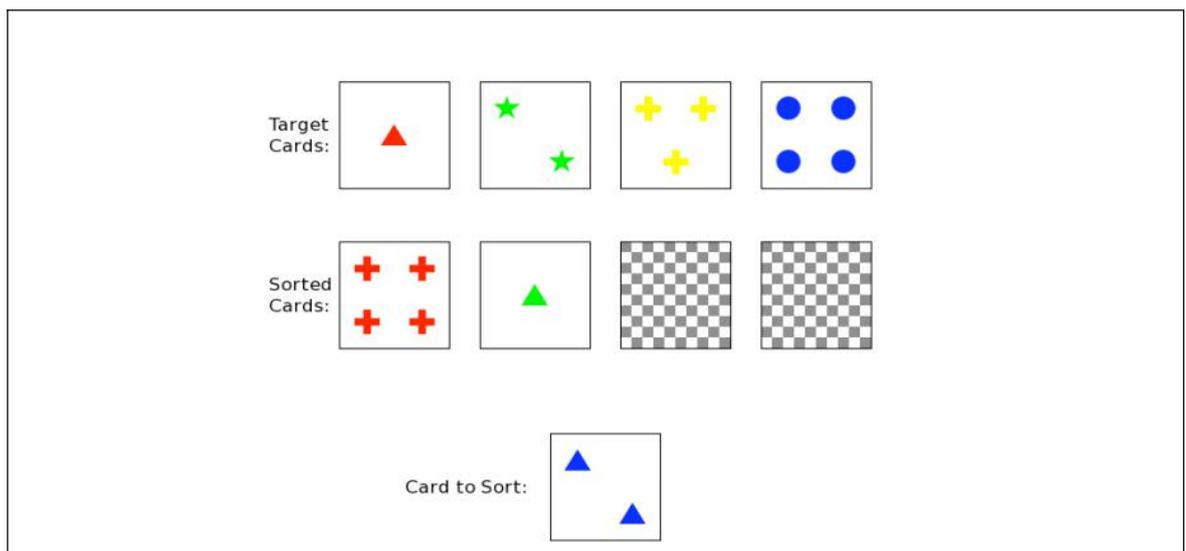


Fig B-1 The Wisconsin Card Sorting Task (WCST). The figure shows the situation after two cards have been sorted according to the proposed schema 'colour' and as a third card (two blue triangles) is presented for sorting.

Cooper, Wulke and Advalaar (2012) conducted this WCST experiment. They explored six dependent measures on the WCST: correct (number of correctly sorted cards), categories (correctly sorted categories), TFC (trials for the first category), CPE (classical perseverative errors), NPE (non-perseverative errors), PP (perseverative proportion). There are some individual differences within these participants' performance. Specifically, as can be seen in Fig

2A, most of the participants ($n = 32$) performed very well (number of sorts > 2). This produced the mean of number of correctly sorted sorts to be 3.77 ($SE = 1.10$).

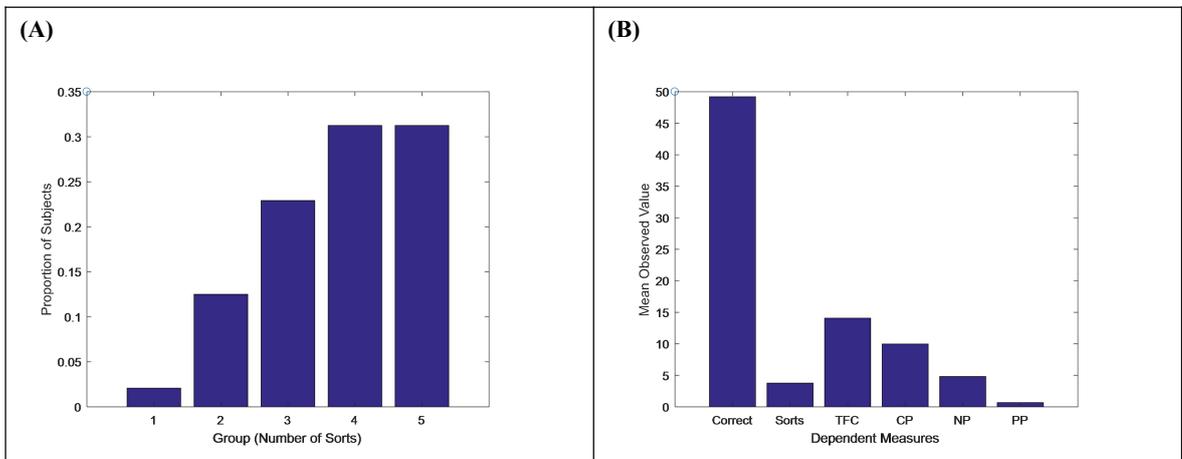


Fig B-2. Observed data from Cooper, Wutke and Davelaar (2012). A: proportion of subjects perform different number of sorts. B: mean observed values of each dependent measure. (TFC = trials for first category; CP = classic perseverative error; NP = nonperseverative error; PP = proportion of perseverative error)

The WCST is a highly set-shifting function involved task. As Cooper et al. (2012) found that, in the concurrent task interference conditions, the WCST performance would be heavily interfered with a set-shifting function-based secondary task. Hence, It requires a model to be able to quickly switch between rules (or cues). However, the attentional learning theory assumes that attention or salience of a cue is obtained by the error-driven learning mechanism.

B.3 The Model

The principles of error-driven attentional learning is implemented in the following model. The attention strengths on input dimensions are sigmoidal functions of contextual bias. The idea stems from Kruschke's (1996) AMBRY model, which was developed to instantiate the theory of dissociable extradimensional and interdimensional shifts in category learning. As mentioned above, the category membership of the item in the attention learning model is determined by its similarity to stored exemplars. However, in the WCST, each card is presented once during the task. It is reasonable to consider that the participants decide the assignment based on the similarity between present card and the target cards. Hence, three input nodes are used to

represent the three dimensions (shape, colour, and number), four exemplar (hidden) nodes are used to represent the target cards, and four response nodes are used to represent the possible responses.

The similarity is assumed to be inversely related to the distance between the perceptual representations of the stimuli. More specifically, the distance between the perceptual representations of stimuli i and exemplar j , denoted α_j^{hidden} , is computed from the weighted Minkowski metric:

$$\alpha_j^{hidden} = \exp(-c \sum_i a_i |I_i - V_{i,j}|) \quad (\text{B.1})$$

where I_i indicates the scale value of the present card on the i th dimension, where c is a constant called the specificity that determines the overall width of the receptive field, a_i is the attention strength on the i th dimension, and where $V_{i,j}$ is the scale value of the j th target card on the i th dimension.

Let the attention strength a_i be a function of some underlying variable β_i , rather than a primitive in the formalisation:

$$a_i = \frac{1}{1 + \exp(-\beta_i)} \quad (\text{B.2})$$

There is one response node per target card, with activation given by:

$$\alpha_k^{out} = \sum_j w_{j,k} \alpha_j^{hidden} \quad (\text{B.3})$$

where $w_{j,k}$ is the association weight to response node k from target card j . The target card-to-response association weights are initialised such that $w_{j,k} = 1$ if $j = k$ and $w_{j,k} = 0$ otherwise. Respond k if $\alpha_k^{out} = \max(\alpha_k^{out})$.

As mentioned earlier, the dimensional attention strengths are learned by gradient descent on error. Each trial is followed by feedback indicating the correct response, just as in the

experiments with human participants. The feedback is coded as ‘humble teacher value’, t_k , given to each response node. For a given trial, the error generated by the model is defined as

$$E = \frac{1}{2} \sum_k (t_k - a_k^{out})^2 \quad (\text{B.4})$$

where the teacher values are defined in these simulations as $t_k = +1$ if the response k is correct, and $t_k = 0$ if the response k is incorrect.

Upon presentation of each trial to the model, the attention strengths are changed so that the error decreases. Following Rumelhart et al. (1986) and Kruschke (1996a), they are adjusted proportionally to the (negative of the) error gradient. Evaluating the derivatives leads to the following delta rules:

$$\Delta\beta_i = -\lambda_a \sum_j \left(\sum_k (t_k - \alpha_k^{out}) w_{j,k} \right) \alpha_j^{hidden} c | I_i - V_{i,j} | a_i (1 - a_i) \quad (\text{B.5})$$

where the λ_a is the constant of proportionality called *attention learning rate*.

Since the target cards are presented across the whole period of task, the implementation of the model does not include the change of target card-to-response association weights.

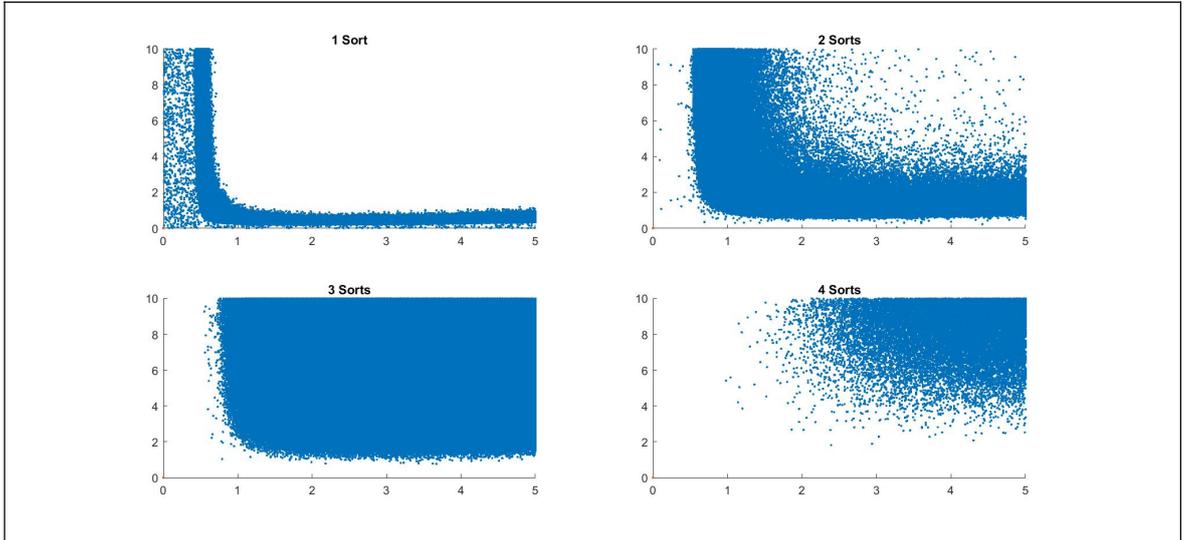


Fig B-3. Parameter values distribution in terms of the number of sort(s) being sorted by the model prediction.

B.4 Results and Discussion

The goal of this article is to figure out the limits of the error-driven attention learning approach, thus we are not going to look at the specific best-fits of this model. First, we look at the interaction between the parameters and the model prediction. Fig 3 shows the parameter values distribution corresponding to the prediction of number of sorts being sorted. The range of variation on parameter c is $[0,5]$ and the range of variation on parameter λ_a is $[0,10]$. The model can mimic the performance of all groups but that of the 5 sorts group.

As can be seen in Fig 3, the model can predict either 3 or 4 sorts performance when $c = 5$ and $\lambda_a = 10$. The simulation ran 300 times. The mean and standard deviation of each dependent measure was computed. As can be seen in Table 1, the model produces much more perseverative errors than human data, and the proportion of perseverative errors is higher than humans. this explains the limits of the attention learning approach that it holds stronger perseveration than humans.

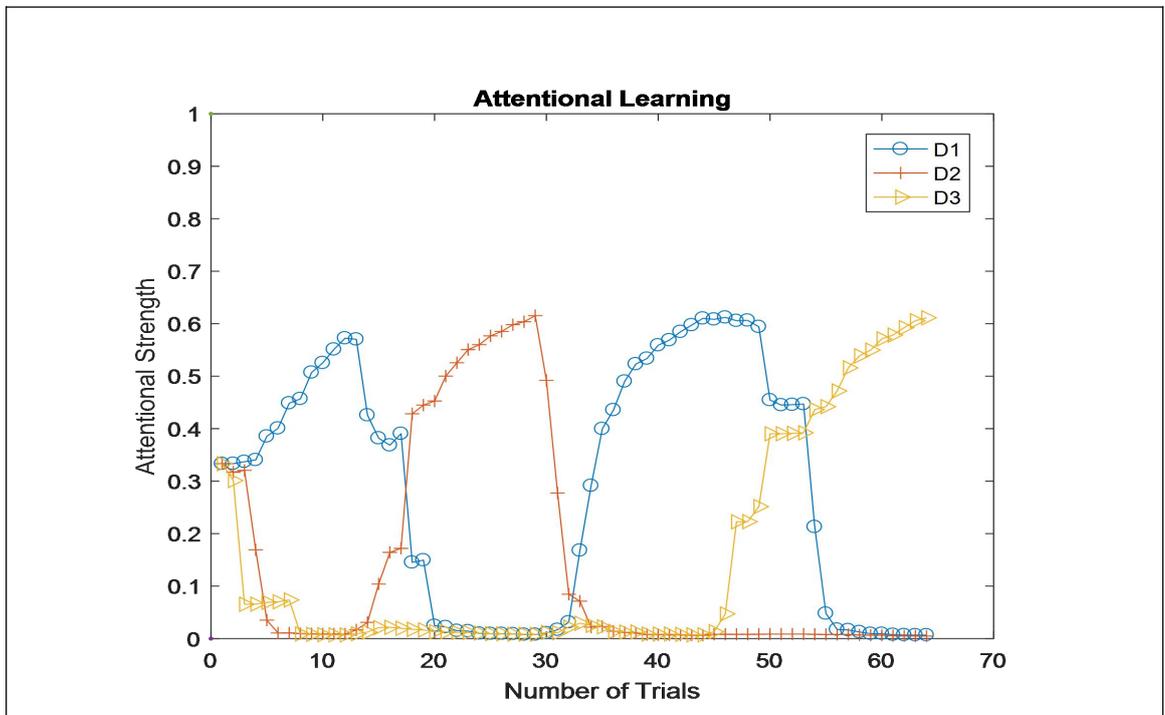


Figure B-4. Results of a simulation of attentional strengths on three dimensions (D1 = shape, D2 = color, D3 = number) during the process of sorting four categories.

As is known that the perseveration is produced by excitation and inhibition in the processes of rule shifting. Fig 4 illustrates an example of the change of attention strengths across the task. It shows that the gradient descent on errors leads to the weakness of the model during the process that requires a rapid excitation and inhibition.

Table B-1
Comparison of Human Data and Model Prediction

| Dependent Measures | Human Data | | Model Prediction | |
|--------------------|--------------|--------|------------------|--------|
| Correct | 49.21 | (6.74) | 46.43 | (2.09) |
| Sort(s) | 3.77 | (1.10) | 3.37 | (0.48) |
| TFC | 14.08 | (7.22) | 12.33 | (1.57) |
| CPE | 10.02 | (4.81) | 16.35 | (2.08) |
| NPE | 4.77 | (2.45) | 1.22 | (0.99) |
| PP | 0.69 | (0.09) | 0.93 | (0.05) |

In sum, the results indicate that the gradient descent error-driven attention is not able to account for the rule shifting in the context requiring quick and frequent switching.

Appendix C.

Modelling Aha and Goldstone (1992)

C.1 Background

Although, ATRIUM has been argued to be advantageous in many empirical phenomena in category learning (e.g., Nosofsky & Little, 2010; Sewell & Lewandowsky, 2011; Yang & Lewandowsky, 2006), to my knowledge, no attempt has been made to examine the capacity of the model to account for some well-established empirical data sets. However, establishing these simulations seems necessarily to be the first step in determining whether the principles of exemplar-based representational attention can provide a good description of the organisation of categorisation behaviour. In this Section, I attempt to demonstrate how the mixture-of-experts approach accounts for a classic heterogeneous categorisation task (Aha & Goldstone, 1992).

A classic study is Aha and Goldstone (1992). Aha and Goldstone (1992) sampled training stimuli from two distinct categories in a two-dimensional stimulus space (see Fig 2-7A), each of which was bisected by its own uniquely oriented boundary. When extended further from their cluster, the boundaries dictated opposite classifications for the same test items. Aha and Goldstone, thus, found that participants classified transfer stimuli using the closest partial boundary (see Fig 2-7B), and this result reveals that selection of categorisation behaviour is dependent on differences between subsets of stimuli.

According to the Ashby et al. (1998) traditional dual systems framework, as explicit rules are controlled by the hypothesis testing system. Shifts between partial rules can be determined by a very complex explicit rule, such as ‘if $y > 5.5$, if $y > 6.5$, then respond B, other wise respond A; if $y < 5.5$, if $x < 6.5$, respond B, otherwise, respond A’. However, this rule may be insufficient to explain the observed data in Aha and Goldstone (1992), because there remains

some area in the two-dimensional space which does not quite fit this rule (see Figure 3-10B). Instead, a number of studies have shown that the exemplar-similarity gated shifts between partial rules can fit the partitioning of task knowledge very well (e.g., Sewell & Lewandowsky, 2011; 2012; Yang & Lewandowsky, 2004). Here, I attempt to illustrate how the gating network accounts for the exemplar-similarity gated rule selection.

C.2 Method

The mixture-of-experts architecture of ATRIUM consists of two types of modules: expert modules and a gating network. Also, as a model based on the multiple representations theory, it consists of two types of expert modules: rule modules and exemplar-based module. Modules independently learn to categorise stimuli and compete to produce output. The gating network adjudicates between the modules, and determines the contribution each makes to the final overall response. Such a modular architecture allows to fit not only dissociations between ERB and II category learning tasks, but also many complex rules. In the original version of ATRIUM,

Table C-1.
Best-fitting parameters for both models for the data from different tasks.

| Parameter | Settings |
|----------------|----------|
| c | 1.0414 |
| ϕ | 3.7146 |
| y_r | 1.0382 |
| λ_{r1} | 0.8373 |
| λ_{r2} | 1.8962 |
| y_g | 1.1570 |
| λ_g | 1.1717 |
| β_{ex} | 1.6754 |
| λ_a | 0.1285 |
| λ_{ex} | 0.3826 |

Note— λ_{r1} represents the rule learning rate of line position rule and λ_{r2} represents the learning rate of the square size rule in Aha and Goldstone Experiment.

As in Aha and Goldstone's (1992) task, it consists of two different partial rules, I has to adapt the rule modules to accommodate the multiple subsets of rules. Thus, the model

implemented here consists of two rule modules. The procedure used to simulate ATRIUM consists of a training phase, which consists of 25 blocks of 12 randomly presented stimuli (i.e., as Aha and Goldstone (1992) did not officially provide training data), and a transfer phase, which consists of 64 randomly presented stimuli, including training stimuli and novel stimuli. Parameter settings used in this simulation is shown in Table C-1.

C.3 Results and Discussion

Fig C-1 shows that ATRIUM did a very job at predicting the proportion of B responses in the transfer phase. In particular, the model successfully categorise item [2, 7] and item [6, 5] (i.e., critical items W and Y) as member of B, and item [6, 5] and item [7, 2] as member of A (i.e., critical items X and Z).

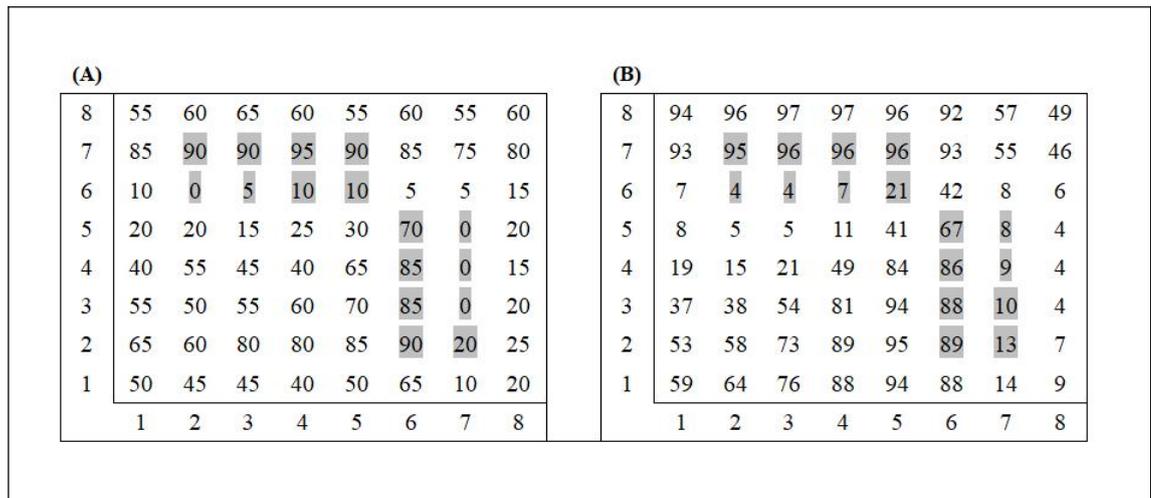


Figure C-1. (A). the proportion of category B responses in the transfer phase, and (B). the proportion of category B responses in the transfer phase predicted by ATRIUM. The horizontal axis denotes line positions and vertical axis denotes square size.

Rather than manipulating a complex rule, ATRIUM fit the transfer data by using exemplar-similarity gated shifts between partial rules. The gating network did a good job in this simulation. One thing must be mentioned here is that the operation of exemplar-similarity gated shifts between partial rules, together with rule-plus-exception category learning, is very similar to the conception of hierarchical organisation of task representation (schema) in the supervisory

system theory (e.g., Cooper & Shallice, 2000, 2006a). As mentioned earlier, representation of a task could consist of multiple lower-level internal representations. Selection of these internal representations is determined by the relationships between individual objects and corresponding schemas. It is indeed analogous to that selection of partial knowledge (e.g., partial rules or exceptions) is driven by the association between each exemplar and the gating network.

References

- Aha, D.W., & Goldstone, R.L. (1992). Concept learning and flexible weighting. In J. K. Kruschke (Ed.), *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.
- Akkal, D., Dum, R.P., & Strick, P.L. (2007). Supplementary motor area and presupplementary motor area: targets of basal ganglia and cerebellar output. *Journal of Neuroscience*, *27*, 10659e10673.
- Alfonso-Reese, L.A., Ashby, F.G., & Brainard, D.H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, *64*(4), 570-583.
- Anderson, J.R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, *8*, 629–647. doi:10.3758/BF03196200.
- Anderson, J.R., Bothwell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*, 1036–1060.
- Aron, A.R., Fletcher, P.C., Bullmore, E.T., Sahakian, B.J., & Robbins, T.W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature: Neuroscience*, *6*(2), 115-116.
- Aron, A.R., Robbins, T.W., & Poldrack, R.A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*(4), 170-177.
- Aron, A.R., Shohamy, D., Clark, J., Myers, C.E., Gluck, M.A., Poldrack, R.A., 2004. Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology*, *92*, 1144–1152.
- Arbuthnott, G.W., Ingham, C.A., & Wickens, J.R. (2000). Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy*, *196*, 587-596.
- Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology*, Volume 2. NY: Cambridge University Press. 223-270.
- Ashby, F.G. & Alfonso-Reese, L.A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*. *39*. 216-233.
- Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*. *105*(3), 442-481;
- Ashby, F.G., & Crossley, M.J. (2010). Interactions between declarative and procedural-learning categorization systems. *Neurobiol. Learn. Mem.* *94*: 1-12;
- Ashby, F.G., & Crossley, M.J. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 363-376.

- Ashby, F.G., & Ell, S.W. (2001). The neurobiology of human category learning. *Towards a Cognitive Science*. 5, 204-210;
- Ashby, F. G., Ell, S. W., Valentin, V., & Casale, M. B. (2005). FROST: a distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, 17, 1728e1743.
- Ashby, F.G., Ell, S.W., & Waldron, E.M. (2003a). Procedural learning in perceptual categorization. *Memory & Cognition*. 31(7). 1114-1125.
- Ashby, F.G., Ennis, J.M., & Spiering, B.J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*. 114(3), 632-656;
- Ashby, F.G., & Gott, R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 14, 33-53;
- Ashby, F.G., & Maddox, W.T. (2010). Human category learning 2.0. *Annals of the New York Academy of Sciences*. 1224, 147-161. Doi:10.1111/j.1749-6632.2010.05874x;
- Ashby, F.B., & Mckinley, W.T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception & Performance*. 16. 598-612;
- Ashby, F.G., Noble. S., Filoteo, J.V., Waldron, E.M., & Ell, S.W. (2003b). Category learning deficits in Parkinson's disease. *Neuropsychology*. 17. 115-124;
- Ashby, F.G., & O'Brien, J.B. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics*. 69(6), 865-878;
- Ashby, F.G., Paul, E.J., & Maddox, W.T. (2010). COVIS. In: E, M. Pothos and A, J. Wills (eds). *Formal Approaches in Categorization*. 65-87. New York: Cambridge University Press.
- Ashby, F.G., & Rosedahl, L. (in press). A neural interpretation of exemplar theory. *Psychological review*.
- Ashby, F.G., & Spiering, B.J. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Review*. 3(2), 101-113;
- Ashby, F. G., Turner, B. O., and Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends Cogn. Sci. (Regul. Ed.)* 14, 191–232.
- Ashby, F.G., & Valentin, V.V. (2016). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen and C. Lefebvre (Eds.), *Categorization in Cognitive Science, 2nd Edition*, New York: Elsevier.
- Ashby, F.G., & Vucovich, L.E. (2016). The role of feedback contingency in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*.
- Ashby, F.G., & Waldron, E.M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*. 6(3). 363-378;
- Badre, D., & Frank, M.J. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 2: Evidence from fMRI. *Cerebral Cortex*. 22. 527-536.
- Balleine, B.W., Lijeholm, M., & Ostlund, S.B. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behav. Brain Res*. 199, 43–52.
- Barbas, H., & Pandya, D.N. (1987). Architecture and frontal cortical connections of the premotor cortex (area 6) in the rhesus monkey. *Journal of Comparative Neurology*, 256, 211-228.

- Beatty, W.W., Staton, R.D., Weir, W.S., Monson, N., & Whitaker, H.A. (1989). Cognitive disturbances in Parkinson's disease. *Journal of Geriatric Psychiatry and Neurology*, 2, 22-33;
- Beldarrain, M.G., Grafman, J., Pascual-Leone, A., & Garcia-Monco, J.C. (1999). Procedural learning is impaired in patients with prefrontal lesions. *Neurology*, 52(9), 1853–1868.
- Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 22, 458-475;
- Blair, M., & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition*, 29(8), 1153-1164;
- Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 35, 1196–1206.
- Bogacz, B., & Cohen, J.D. (2004). Parameterization of connectionist models. *Behavior research methods, Instruments & Computers*, 36(4), 732-741;
- Bolam, J.P., Powell, J.F., Wu, J.Y., & Smith, A.D. (1985). Glutamate decarboxylase-immunoreactive structures in the rat neostriatum: a correlated light and electron microscopic study including a combination of Golgi impregnation with immunocytochemistry. *Journal of Comparative Neurology*, 237, 1-20.
- Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society*, B369: 20130480;
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652.
- Braver, T.S. (2012). The variable nature of cognitive control: A dual-mechanisms framework. *Trends in Cognitive Sciences*, 16(2), 106-113. doi:10.1016/j.tics.2011.12.010;
- Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, 55(1), 37–85.
- Chandrasekaran, B., Koslov, S.R., & Maddox, W.T. (2014). Toward a dual-learning systems model of speech category learning. *Front Psychol*, 5, 825, doi: 10.3389/psvg.2014.00825;
- Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 351(1346), 1515–1527.
- Cohen, M.X., & Frank, M.J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav. Brain Res*, 199, 141–156.
- Collins, A.G.E., Brown, J.K., Gold, J.M., Waltz, J.A., & Frank, M.J. (2014). Working memory contribution to reinforcement learning impairments in schizophrenia. *The Journal of Neuroscience*, 34(41), 13747-13756.
- Collins, A.G.E., & Frank, M.J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35, 1024-1035.

- Collins, A.G.E., & Frank, M.J. (2013). Cognitive control over learning: Clustering, and generalization task-set structure. *Psychological Review*, *130*(1), 190-229;
- Contwell, G., Crossley, M.J., & Ashby, F.G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, 1-16.
- Cooper, R.P. (2010a). Cognitive control: Componential or emergent? *Topics in Cognitive Science*, *2*(4), 598-613.
- Cooper, R.P. (2010b). Forward and inverse models in motor control and cognitive control. In J. Chappell, S. Thorpe, N. Hawes, & A. Sloman (eds), *Proceedings of the Symposium on AI-Inspired Biology (AIIB)*, (part of AISB 2010), Leicester, UK.
- Cooper, R.P. (2011). Complementary perspectives on cognitive control. *Topics in Cognitive Science*, *3*(2), 208-211.
- Cooper, R.P., & Marsh, V. (2015). Set-shifting as a component process of goal-directed problem solving. *Psychological Research*. Doi.10.1007/s00426-015-0652-2;
- Cooper, R.P., Ruh, N., & Mareschal, D. (2014). The Goal Circuit Model: A hierarchical multi-route model of the acquisition and control of routine sequential action in humans. *Cognitive Science*, *38*, 244-274. DOI: 10.1111/cogs.12067.
- Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297–338.
- Cooper, R.P., & Shallice, T. (2006a). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887-916.
- Cooper, R.P., & Shallice, T. (2006b). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297-338.
- Cooper, R. P., Shallice, T., & Farrington, J. (1995). Symbolic and continuous processes in the automatic selection of actions. In J. Hallam (Ed.), *Hybrid problems, hybrid solutions* (pp. 27–37). Amsterdam: IOS Press.
- Cooper, R.P., Wutke, K., & Davelaar, E.J. (2012). Differential contributions of set-shifting and monitoring to dual-task interference. *The Quarterly Journal of Experimental Psychology*, *63* (3), 587-612.
- Corter, J., & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*, 291–303.
- Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you to learn. *The Quarterly Journal of Experimental Psychology*, *65*(3), 439-464.
- Craig, S., & Lewandowsky, S. (2013). Working memory supports inference learning just like classification learning. *The Quarterly Journal of Experimental Psychology*, *66*(88), 1493-1403.
- Crossley, M.J., & Ashby, F.G. (2015). Procedural Learning during Declarative Control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1388-1403.
- Crossley, M.J., Ashby, F.G., & Maddox, W.T. (2012). Enrasing the engram: The unlearning of procedural skills. *Journal of Experimental Psychology: General*, *142*(3), 710-741;
- Crossley, M.J., Paul, E.J., Roeder, J.L., Ashby, F.G. (2015). Declarative strategies persist under increased cognitive load. *Psychon Bull Rev*. Doi: 10.3758/s13423-015-0867-7

- Davidson, M.C., Amso, D., Anderson, L.C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037-2078;
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature: Neuroscience*, *8*, 1704–1711.
- Dehaene, S. & Changeux, J. P. (1997). A hierarchical neuronal network for planning behaviour. *Proceedings of the National Academy of Science, USA*, *94*, 13293-13298.
- Denton, S.E., Kruschke, J.K., & Erickson, M.A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin & Review*, *15*(4), 780-786;
- Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*, 732-739.
- Dum, R.P., & Strick, P.L. (1991). The origin of corticospinal projections from the premotor areas in the frontal lobe. *Journal of Neuroscience*, *11*, 667-689.
- Duncan, J. (1993). Selection of input and goal in the control of behaviour. In D. A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control—A tribute to Donald Broadbent* (pp. 53–71). Oxford, England: Oxford University Press.
- Eimas, P.D., & Quinn, P.C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, *65*, 903–917.
- Ell, S.W., Ashby, F.G., & Hutchinson, S. (2012). Unsupervised category learning with integral dimension stimuli. *The Quarterly Journal of Experimental Psychology*, *65*(8), 1537-1562. Doi: [10.1080/17470218.2012.658821](https://doi.org/10.1080/17470218.2012.658821).
- Ell, S.W., Marchant, N.L., & Ivry, R.B. (2006). Focal putamen lesions impair learning in rule-based, but not informationintegration categorization tasks. *Neuropsychologia*, *44*(10), 1737–1751.
- Erickson, M.A. (2008). Executive attention and task-switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*(4), 749-761;
- Erickson, M.A., & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107-140;
- Erickson, M.A., & Kruschke, J.K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic, Bulletin & Review*, *9*(1), 160-168;
- Feenstra, M.G., & Botterblom, M.H. (1996). Rapid sampling of extracellular dopamine in the rat prefrontal cortex during food consumption, handling and exposure to novelty. *Brain Research*, *742*, 17-24.
- Fernandez-Ruiz, J., Wang, J., Aigner, T. G., & Mishkin, M. (2001). Visual habit formation in monkeys with neurotoxic lesions of the ventrocaudal neostriatum. *Proceedings of the National Academy of Sciences*, *98*, 4196–4201.
- Filoteo, J.V. & Maddox, W.T. (2007). Category learning in Parkinson’s disease. In *Research progress in Alzheimer’s disease and dementia*. Maio-Kun Sun, Ed.: Vol. 3, 2–26. Nova Science Publishers, Inc. Hauppauge, NY.
- Filoteo, J.V., & Maddox, W.T., & Davis, J.D. (2001). A possible role of striatum in linear and nonlinear category learning: Evidence from patients with Huntington’s disease. *Behavioral neuroscience*, *115*(4), 786-798;
- Filoteo, J.V., Maddox, W.T., Salmon, D.P., & Song, D.D. (2005). Information integration category learning in patients with striatal dysfunction. *Neuropsychology*, *19*, 212-222;

- Fisher, A.V., & Slousky, V.M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development*, *76*, 583-597;
- Frank, M.J. (2005). Dynamic dopamine modulation in the basal ganglia: a neu-rocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *J. Cogn. Neurosci.* *17*, 51–72.
- Frank, M.J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*. *22*. 509-526.
- Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*, 137–160.
- Fuster, J.M. (1989). *The Prefrontal Cortex*. New York: Raven.
- Fuster, J.M. (2008). *The prefrontal cortex* (4th ed.). Singapore, CN: Academic Press.
- Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, *4*, 161–178.
- George, D.N., & Kruschke, J.K. (2012). Contextual modulation of attention in human category learning. *Learn Behav.* *40*: 530-541. Doi: 10.3758/s13420-012-0072-8.
- Gershman, S.J., & Blei, D.M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*. *56*. 1-12.
- Gilbert, S. J., & Shallice, T. (2002). Task switching: a PDP model. *Cogn. Psychol.* *44*, 297–337.
- Gluck, M.A., & Bower, G.H. (1988). from conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227-247.
- Gluck, M.A., Shohamy, D., & Myers, C. (2002). How do people solve the ‘weather prediction’ task: individual variability in strategies for probabilistic category learning. *Learning & Memory*. *9*. 408-418;
- Gotham, A.M., Brown, R.G., & Marsden, C.D. (1988). ‘Frontal’ cognitive function in patients with Parkinson’s disease.’on’ and ‘off’ levodopa. *Brain*, *111*, 299-321.
- Graf, P., and Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *J. Exp. Psychol. Learn. Mem. Cogn.* *11*, 501–518.
- Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* *31*, 359–387.
- Helie, S., Ell, S.W., & Ashby, F.G. (2016). Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex*. *64*, 123-135.
- Helie, S., Paul, E.J., & Ashby, F.G. (2012a). A neurocomputational account of cognitive deficits of Parkinson’s disease. *Neuropsychologia*. *50*, 2290-2302.
- Helie, S., Paul, E.J., & Ashby, F.G. (2012b). Simulating the effects of dopamine imbalance on cognition: From positive effect to Parkinson’s disease. *Neural Networks*. *32*. 74-85.
- Hélie, S., Roeder, J.L., & Ashby, F.G., (2010a). Evidence for cortical automaticity in rule-based categorization. *J. Neurosci.* *30*, 14225–14234.
- Helie, S., Roeder, J.L., Vucovich, L., Runger, D., & Ashby, F.G. (2015). A neurocomputational model of automatic sequence prediction. *Journal of Cognitive Neuroscience*. *27*(7). 1456-1469. Doi: 10.1162/jocn_a_00794.

- Helie, S., Waldschmidt, J.G., & Ashby, F.G. (2010b). Automaticity in rule-based and information integration category learning. *Attention, Perception, & Psychophysics*, *72*(4), 1013-1031.
- Helversen, B., Karlsson, L., Rasch, B., & Rieskamp, J. (2014). Neural substrates of similarity and rule-based strategies in judgment. *Front Psychol*, *8*, 809. doi: 10.3388/nhum.2014.00809;
- Hoffman, A.B., & Rohder, B. (2010). The costs of supervised classification: the effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319-340;
- Homa, D. (1984). On the nature of categories. *Psychology of Learning & Motivation*, *18*, 49-94.
- Jenkins, I. H., Brooks, D. J., Nixon, P. D., Frackowiak, R. S., & Passingham, R. E. (1994). Motor sequence learning: A study with positron emission tomography. *Journal of Neuroscience*, *14*, 3775–3790.
- Johansen, M.K., Fouquet, N., & Shanks, D.R. (2007). Paradoxical effects of base rates and representation in category learning. *Memory & Cognition*, *35*(6): 1365-1379.
- Johansen, M.K., & Kruschke, J.K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1433-1458.
- Jueptner, M., Stephan, K.M., Frith, C.D., Brooks, D.J., Frackowiak, R.S.J., & Passingham, R.E. (1997). Anatomy of motor learning. 1. Frontal cortex and attention to action. *Journal of Neurophysiology*, *77*, 1313–1324.
- Kalish, M.L., & Kruschke, J.K. (1997). Decision boundaries in one-dimensional categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*, 1362-1377;
- Katahira, Y.B., & Ohira, H. (2014). Dual-learning processes underlying human decision-making in reversal learning tasks: Functional significance and evidence from the model fit to human behavior. *Front Psychol*, *5*, 871. doi: 10.3389/psvg.2014.00871;
- Kemp, J.M., & Powell, T.P. (1971). The structure of the caudate nucleus of the cat: light and electron microscopy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *262*, 383-401.
- Kieffaber, P.D., Kruschke, J.K., Cho, R.Y., Walker, P.M., & Hetrick, W.P. (2013). Dissociating stimulus-set and response-set in the context of task-set switching. *Journal of Experimental Psychology: Human Perception & Performance*, *39*(3), 700-719;
- Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—a review. *Psychological Bulletin*, *136*(5), 849–874.
- Kloos, H., & Slousky, V.M. (2008). What’s behind kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*, 52-72;
- Knowlton, B.J., Mangels, J.A., & Squire, L.R. (1996). A neuro-striatal habit learning system in humans. *Science*, *273*, 1399-1402;
- Knowlton, B.J., & Squire, L.R. (1993). The learning of categories: Parallel brain systems for item memory and category level knowledge. *Science*, *262*, 1747-1749;
- Knowlton, B.J., Squire, L.R., & Gluck, M.A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*, 106-120;

- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44;
- Kruschke, J.K. (1993). Human category learning: Implications for back propagation models. *Connection Science*, *5*, 3–36.
- Kruschke, J.K. (1996a). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–247.
- Kruschke, J.K. (1996b). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1): 3-26.
- Kruschke, J.K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812-863.
- Kruschke, J.K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *34*(3), 210-226;
- Kruschke, J.K. (2011). Models of attentional learning. In: E, M. Pothos and A, J. Wills (eds). *Formal Approaches in Category Learning*. 120-152. Cambridge University Press;
- Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*, 636-645.
- Kruschke, J. K., & Erickson, M. A. (1994). Learning of roles that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514--519). Hillsdale, NJ: Erlbaum.
- Kruschke, J.K., & George, D.N. (2012). Contextual modulation of attention in human category learning. *Learn Behav*, *40*, 530-541;
- Kruschke, J.K., & Johansen, M.K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(5), 1083–1119.
- Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.
- Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*(2), 227–260.
- Lamberts, K., & Kent, C. (2007). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition*, *35*, 2097–2105.
- Lavie, N., Hirst, A., Fockert, J.W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*(3), 339-354;
- Lee, M.D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities, and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive science*, *32*, 1403-1424;
- Le Pelley, M.E., HaselGrove, M., & Esher, G.R. (2012). Modeling attention in associative learning: Two processes or one? *Learn Behav*, *40*, 292-304;
- Lewandowsky, S. (2011). Working memory capacity and categorization: Individual difference and modeling. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *37*(3): 720-738.
- Lewandowsky, S., Roberts, L., & Yang, L. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition*, *34*(8), 1676-1688;

- Little, D.R., & Lewandowsky, S. (2009). Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception & Performance*, *35*(2), 530-550;
- Little, D.M., & Thulborn, K.R. (2006). Prototype-distortion category learning: A two-phase learning process across a distributed network. *Brain and Cognition*, *60*, 233-243;
- Lombardi, W. J., Andreason, P. J., Sirocco, K. Y., Rio, D. E., Gross, R. E., Umhau, J. C., & Hommer, D. W. (1999). Wisconsin Card Sorting Test performance following head injury: Dorsolateral fronto-striatal circuit activity predicts perseveration. *Journal of Clinical and Experimental Neuropsychology*, *21*, 2-16.
- Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309-332;
- Luria, A.R. (1966). *Higher Cortical Functions in Man*. London: Tavistock .
- Mack, M.L., & Palmeri, T.J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General*, *144*(3), 551-569.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163-203.
- Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49-70.
- Maddox, W.T., Ashby, F.G. & Bohil, C.J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *J. Exp. Psychol.: Learn., Mem., Cognit.* **29**: 650-662.
- Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004a). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, **11**, 945-952.
- Maddox, W.T., & Filoteo, J.V. (2001). Striatal contributions to category learning: Quantitative modeling of simple linear and complex nonlinear rule learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society*, *7*, 710-727;
- Maddox, W.T., Filoteo, J.V., Hejl, K.D., & Ing, A.D. (2004b). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category learning systems. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*(1), 227-245.
- Maddox, W.T., Glass, B.D., O'Brien, J.B., Filoteo, J.B., & Ashby, F.G. (2010a). Category label and response location shifts in category learning. *Psychological Research*, *74*: 219-236.
- Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100-107.
- Maddox, W.T., Lauritzen, S.J., & Ing, A.D. (2007). Cognitive complexity effects in perceptual classification are dissociable. *Memory & Cognition*, *35*, 885-894.
- Maddox, W.T., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D.M. (2010b). Rule-based and information-integration category learning in normal aging. *Neuropsychologia*, doi: 10.1016/j.neuropsychologia.2010.06.008;
- Markman, A.B., & Ross, B.H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592-613.

- McClelland, J. L. (1992). Toward a theory of information processing in graded, random, interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp 655– 688). Cambridge, MA: MIT Press.
- McDaniel, M.A., Cahill, M.J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*. 143(2). 668-693.
- Mckinley, S.C., & Nosofsky, R.M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, 21, 128-148;
- Mckinley, W.T., & Nosofsky, R.M. (1996). Selective attention and the function of linear decision boundaries. *Journal of Experimental Psychology: Human Perception & Performance*. 22. 294-317;
- Medin, D.L., Alton, M.W., Edelson, S.M., & Freko, D. (1982). Correlated symptoms and simulated medial classification. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 8, 37-50;
- Medin, D.L., Dewey, G.I., & Murphy, T.D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 9, 607-625;
- Medin, D.L., Lynch, E.B., Coley, J.D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, 32, 49–96.
- Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238;
- Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*. 7, 355-368;
- Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1423-1442
- Miller, E.K., & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167-202.
- Milton F1, & Pothos EM. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*. 34. 1326-1336.
- Miskin, M., Malamut, B., and Bachevalier, J. (1984). “Memories and habits: two neural systems,” in *Neurobiology of Learning and Memory*, eds. G. Lynch, J. L. McGaugh, and N. M. Weinberg (New York: Guilford), 65–67.
- Miyake, A., Fredman, N.P., Emerson, M.J., Witrki, A.H., Hovetter, A., & Wager, T.D. (2000). The utility and diversity of executive functions and their contributions to complex ‘frontal lobe’ tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49-100.
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134–140.
- Montague, P.R., Dayan, P., & Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16, 1936-1947.
- Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews: Neuroscience*, 9, 856-869.

- Nakano, K., Kayahara, T., Tsutsumi, T., and Ushiro, H. (2000). Neural circuits and functional organization of the striatum. *J. Neurol.* 247, V1–V15.
- Neil, G.J., & Higham, P.A. (2012). Implicit learning of conjunctive rule sets: an alternative to artificial grammars. *Consciousness & Cognition.* 21, 1393-1400;
- Newell, B. R., Dunn, J. C., & Kalish, M. L. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition,* 38, 563–581.
- Norman, D. A., & Shallice, T. (1980). *Attention to action: Willed and automatic control of behavior* (Chip Rep. No. 99). San Diego: University of California.
- Norman, D.A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. Davidson, G. Schwartz, & D. Shapiro (Eds.), *Consciousness and self regulation*, Vol. 4 (pp. 1–18). New York: Plenum.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General.* 115(1), 39-57.
- Nosofsky, R.M. (2011). The generalized context model: an exemplar model of classification. In: E. M. Pothos and A. J. Wills (eds). *Formal Approaches in Categorization.* 18-39. New York: Cambridge University Press.
- Nosofsky, R.M., Grack, M.A., Palmeri, T.J., Mckinley, S.C., & Glauthier, P. (1994a). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition.* 22, 352-369.
- Nosofsky, R.M., & Johansen, M. (2000). Exemplar-based accounts of ‘multiple-system’ phenomena in perceptual categorization. *Psychonomic Bulletin & Review.* 7(3), 373-402.
- Nosofsky, R.M., & Little, D.R. (2010). Classification response times in probabilistic rule-based category structures: Contrasting exemplar-retrieval and decision-boundary models. *Memory & Cognition.* 38(7): 916-927;
- Nosofsky, R.M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review,* 104(2), 266-300.
- Nosofsky, R.M., & Palmeri, T.J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review.* 5(3), 345-369.
- Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994b). Rule-plus-exception model of classification learning. *Psychological Review.* 101, 53-79.
- Nosofsky, E.M., Stanton, R.D., & Zaki, S.R. (2005). Procedural interference in perceptual classification: implicit learning or cognitive complexity. *Memory & Cognition.* 33(7). 1256-1271;
- O'Reilly, R.C., & Frank, M.J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation.* 18. 283-328.
- Otto, A.R., Skatova, A., Madlon-Kay, S., & Daw, N.D. (2014). Cognitive control predicts use of model-based reinforcement. *Journal of Cognitive Neuroscience,* 27(2), 319-333, doi: 10.1162/jocn_a_00709;
- Palmeri, T.J. (1997). Exemplar similarity and development of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 23(2). 324-354.
- Passingham, R.E., Rowe, J.B., & Sakai, K. (2005). Prefrontal cortex and attention to action. In G. Humphreys & M. J. Riddoch (Eds.), *Attention in action* (pp. 263–286) Hove, UK: Psychology Press.

- Paul, E.J., & Ashby, F.G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*, 7, 177.
- Paul, E.J., Boomer, J., Smith, J.D., Ashby, F.G. (2011). Information-integration category learning and the human uncertainty response. *Memory & Cognition*, 39: 536-554;
- Paul, E.J., Smith, J.D., Valentin, V.V., Turner, B.O., Barbey, A.K., & Ashby, F.G. (2015). Neural networks underlying the metacognitive uncertainty response. *Cortex*, 71, 306-322.
- Pothos, E.M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, 107(2), 581–602;
- Price, A., J.V. Filoteo & W.T. Maddox. 2009. Rule-based category learning in patients with Parkinson’s disease. *Neuropsychologia* 47: 1213–1226.
- Pycock, C.J., Kerwin, R.W., & Carter, C.J. (1980). Effect of lesion of cortical dopamine terminals on subcortical dopamine receptors in rats. *Nature*, 286:74-66.
- Quinn, P.C., Eimas, P.D., & Rosenkrantz, S.L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22, 463–475.
- Rao, S.M., Bobholz, J.A., Hammeke, T.A., Rosen, A.C., Woodley, S.J., Cunningham, J.M., Cox, R.W., Stein, E.A., & Binder, J.R. (1997). Functional MRI evidence for subcortical participation in conceptual reasoning skills. *NeuroReport*, 8: 1987–1993.
- Reder, L.M., Park, H., & Kieffaber, P.D. (2009). Memory systems do not divide on consciousness: Reinterpreting memory in terms of activation and binding. *Psychological Bulletin*, 135(1), 23-49;
- Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M.R., and Obeso, J.A. (2010). Goal-directed and habitual control in the basal gan-glia: implications for Parkinson’s dis-ease. *Nat. Rev. Neurosci.* 11, 760–772.
- Reetzke, R., Maddox, W.T., Chandrasekaran, B. (2016). The role of age and executive function in auditory category learning. *Journal of Experimental Child Psychology*. 142. 48-65.
- Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, 12, 142–162.
- Richer, F., Chouinard, M.J., & Rouleau, I. (1999). Frontal lesions impair the attentional control of movements during motor learning. *Neuropsychologia*, 37, 1427–1435.
- Robinson, C.M., & Sloutsky, V.M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, 11(3), 233-253;
- Rodd, J.M., Johnsrude, I.S., & Davis, M.H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. *Brain & Language*, doi: 10.1016/j.bandl.2010.07.005;
- Roeder, J.L., & Ashby, F.G. (2016). What is automatized during perceptual categorization? *Cognition*.
- Roeder, J.L., Maddox, W.T., & Filoteo, J.V. (2016). The neuropsychology of perceptual category learning. To appear in H. Cohen & C. Lefebvre (Ed.) *Handbook of Categorization in Cognitive Science*, second edition. Elsevier, Ltd.
- Rouder, J.N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, 133, 63-82;

- Rodrigues, P.M., & Murre, J.M.J. (2007). Roles-plus-exception tasks: A problem for exemplar model. *Psychonomic Bulletin, & Review*, 14(4), 640-646;
- Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, 12, 142–162.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2), 207-231.
- Roy, J.E., Riesenhuber, M., Poggio, T., & Miller, E.K. (2010). Prefrontal cortex activity during flexible categorization. *The Journal of Neuroscience*, 30(25), 8519-8528.
- Salminen, N.H., Tittinen, H., & May, P.J.C. (2009). Modeling the categorical perception of speech sounds: A step toward biological plausibility. *Cognitive, Affective, & Behavioral Neuroscience*, 9(3), 304-313;
- Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain. *Neuron*, 20, 185–195.
- Schacter, D. L., & Graf, P. (1986). Effects of elaborative processing on implicit and explicit memory for new associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 432–444.
- Schnyer, D.M., Maddox, W.T., Ell, S., Davis, S., Pacheco, J. & Verfaellie, M, (2009). Prefrontal contributions to rule-based and information integration category learning. *Neuropsychologia*, 47, 2995-3006.
- Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.
- Scimeca, J.M., & Badre, D. (2012). Striatal contribution to declarative memory retrieval. *Neuron*, 75, 380-392;
- Seger, C.A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci. Biobehav. Rev.* 32, 265–278.
- Seger, C.A., & Spiering, B.J. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, 1-9. Doi: 10.3389/fnsys.2011.00066
- Sewell, D.K., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology*, 62, 81–122.
- Sewell, D.K., & Lewandowsky, S. (2012). Attention and working memory capacity: Insights from blocking, highlighting, and knowledge restructuring. *Journal of Experimental Psychology: General*, 141(3), 444-469.
- Sexton, N. & Cooper, R.P. (2015). The role of conflict in the n-2 repetition cost in task switching: a computational model. In D.C. Noelle, R. Dale, A.S. Warlaumont., J. Yoshimi, T. Matlock, C.D. Jennings & P.P. Maglio (eds), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2134-2139). Cognitive Science Society Incorporated, Pasadena, CA.
- Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society, London B*, 298, 199-209.
- Shallice, T. (2002). Fractionation of the supervisory system. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 261–277). Oxford, UK: Oxford University Press
- Shallice, T. (2006). Contrasting domains in the control of action: The routine and the non-routine. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and*

- cognitive development. Attention and performance XXI (pp. 3–29). Oxford, UK: University Press.
- Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 351(1346), 1405–1412.
- Shallice, T., & Cooper, R. P. (2011). *The organisation of mind*. Oxford, UK: Oxford University Press.
- Shallice, T., Stuss, D.T., Picton, T.W., Alexander, M.P. & Gillingham, S. (2008). Mapping task switching in frontal cortex through neuropsychological group studies. *Frontiers in Neuroscience*, 2(1), 79-85.
- Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classification. *Psychological Monographs*. 75,517.
- Shiffrin, W., & Schneider, R.M. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychol. Rev.* 84, 1–66.
- Sloutsky, V.M. (2010). From perceptual categories to concepts: What develops? *Cognitive science*. 34, 1244-1286;
- Sloutsky, V.M., & Fisher, A.V. (2008). Attentional learning and flexible induction: How mundane mechanisms give rise to smart behaviors. *Child Development*, 79, 639-651;
- Sloutsky, V.M., & Robinson, C.W. (2008). The role of words and sounds in infants' visual processing: From overshadowing to attentional tuning. *Cognitive Science*, 32, 342-365;
- Smith, J.D, et al. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience and Biobehavioral Review*. <http://dx.dot.org/10.1016/j.neubiorev.2012.09.003>;
- Smith, J.D., & Minda, J.P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 3–27.
- Smith, J.D., Zakrzewski, A.C., Herberger, E.R., Boomer, J., Roeder, J.L., Ashby, F.G., Church, B.A. (2015). The time course of explicit and implicit categorization. *Atten Percept Psychophys*. 77. 2476-2490.
- Sood, M. & Cooper, R.P. (2013). Modelling the Supervisory System and frontal dysfunction: An architecturally grounded model of the Wisconsin Card Sorting Task. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (eds), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1354-1359). Cognitive Science Society Incorporated, Berlin, Germany.
- Soto, F.A., & Ashby, F.G. (2015) Categorization training increases the perceptual separability of novel dimensions. *Cognition*, 139, 105-129.
- Soto, F. A., Waldschmidt, J. G., Helie, S., & Ashby, F. G. (2013). Brain activity across the development of automatic categorization: A comparison of categorization tasks using multi-voxel pattern analysis. *Neuroimage*, 71, 284–897.
- Spiering, B.J., & Ashby, F.G. (2008). Response processes in information integration category learning. *Neurobiology of Learning and Memory*. 90: 330-338.
- Squire L. R. (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review*. 99(2):195-231.
- Squire, L.R., & Zola-Morgan, S. (1988). Memory: brain systems and behavior. *Trends Neurosci*. 11, 170–175.

- Squire, L.R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science* 253, 1380–1386.
- Stocco, A., Lebiere, C., and Anderson, J.R. (2010). Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychol. Rev.* 117, 541–574.
- Teng, E., Stefanacci, L., Squire, L. R., & Zola, S. M. (2000). Contrasting effects on discrimination learning after hippocampal lesions and conjoint hippocampal-caudate lesions in monkeys. *Journal of Neuroscience*, 20, 3853–3863.
- Thomas, R.D. (1998). Learning correlation in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 24, 119-143;
- Tulving, E., & Schacter, D. L. (1990, January 19). Priming and human memory systems. *Science*, 247, 301–306.
- Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, 108, 550–592.
- Valentin, V.V., Dickinson, A., & O'Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* 27, 4019–4026.
- Valentin, V.V., Maddox, W.T., & Ashby, F.G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. *Frontiers in Psychology*, 5, 643.
- Waldron, E.M., & Ashby, F.G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, 8, 168–176.
- Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *Neuroimage*, 56(3), 1791-1802.
- Westermann, G., Mareschal, D., & Newport, M. (2014). From perceptual to language-mediated categorization. *Phil, Trans, R, Soc, B*, 369: 20120391;
- Wills, A.J., Noury, M., Moberly, N.J. (2006). Formation of category representation. *Memory & Cognition*. 34(1): 17-27;
- Yamauchi, T., & Markman, A.B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124–148.
- Yamauchi, T., & Markman, A. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. 26(3), 776-795;
- Yamauchi, T., & Yu, N. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition*, 36(3), 544-553;
- Yang, L., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 663–679.
- Yang, L., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 30(5). 1045-1064;
- Yang, L., & Wu, Y. (2014). Category variability effect in category learning with auditory stimuli. *Front Psychol*, 2(5), 1122. doi: 10.3389/fpsyg.2014.01122
- Yin, H.H., & Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7, 464-476.

- Younger, B.A., & Cohen, L.B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development*, *57*, 803–815.
- Zetthamova, D., & Maddox, W.T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*. *34*(2), 387-398;