

Systems of category learning: fact or fantasy?

Ben R. Newell<sup>1</sup>

John C. Dunn<sup>2</sup>

Michael Kalish<sup>3</sup>

1. School of Psychology, University of New South Wales, Sydney, Australia

2. School of Psychology, University of Adelaide, Adelaide, Australia.

3. Institute of Cognitive Science, University of Louisiana at Lafayette, USA.

Address for correspondence:

Ben R. Newell

School of Psychology

University of New South Wales

Sydney 2052

Australia

Tel: +61 2 9385 1606

Fax: +61 2 9385 3641

ben.newell@unsw.edu.au

\*\*In press in *The Psychology of Learning & Motivation, Vol 54 (Ed. Brian H. Ross)\*\**

## Table of Contents

### Abstract

### 1. Introduction

#### 1.1 Multiple systems of category learning

### 2. Review and Critique of the Evidence I: Probabilistic Category Learning

#### 2.1 Neuropsychological dissociations

#### 2.2 Re-evaluating the neuropsychological evidence

#### 2.3 Behavioral dissociations

#### 2.4 Re-considering behavioral dissociations

#### 2.5 Neuroimaging

#### 2.6 Re-imagining neuroimaging

#### 2.7 Section Summary

### 3. Review and Critique of the Evidence II: Deterministic Category Learning

#### 3.1 Neuropsychological dissociations

#### 3.2 Re-evaluating the neuropsychological evidence

#### 3.3 Behavioral dissociations

#### 3.4 Re-considering behavioral dissociations

#### 3.5 Neuroimaging

#### 3.6 Re-imagining neuroimaging

#### 3.7 Section Summary

### 4. Re-examining some Fundamental Assumptions

#### 4.1 State-trace analysis

#### 4.2 The inferential limits of dissociations

4.3 A state-trace re-analysis of behavioral and other dissociations

4.4 Interpretation of two-dimensional state-trace plots: The role of confounds

4.5 Interpretation of two-dimensional state-trace plots: Systems, processes and latent variables.

## 5. The Contribution of Mathematical Modelling

5.1 Mathematical/computational models of category learning

5.2 COVIS

5.3 The nature and use of formal models

5.4 Model comparisons

## 6. Discussion and Conclusions

6.1 Varieties of explanation

6.2 Explanatory power

6.3 Final thoughts

## Abstract

Psychology abounds with vigorous debates about the need for one or more underlying mental processes or systems to explain empirical observations. The field of category learning provides an excellent exemplar. We present a critical examination of this field focussing on empirical, methodological and mathematical modelling considerations. We review what is often presented as the ‘best evidence’ for multiple systems of category learning and critique the evidence by considering three questions: (1) Are multiple-systems accounts the only viable explanations for reported effects? (2) Are the inferences sound logically and methodologically? (3) Are the mathematical models that can account for behavior sufficiently constrained, and are alternative (single-system) models applicable? We conclude that the evidence for multiple-systems accounts of category learning does not withstand such scrutiny. We end by discussing the varieties of explanation that can be offered for psychological phenomena and highlight why multiple-systems accounts often provide an illusory sense of scientific progress.

## 1. INTRODUCTION

Psychology has seen an upsurge in theories and accounts that use multiple, qualitatively distinct systems to explain empirical phenomena. Researchers in the areas of judgment and decision making, reasoning, social cognition, associative learning, and memory to name a few, have been swept up in a desire to describe and group behaviors as the products of different systems (see Evans, 2008; Keren & Schul, 2009; Mitchell, DeHouwer, & Lovibond, 2009 for relevant discussions). This desire seems to have been exacerbated in recent years by the introduction of increasingly sophisticated brain imaging methods that allow us to ‘see’ which parts of the brain are ‘recruited’ by different tasks. The engagement of neuroanatomically distinct regions of the brain seems to compel the conclusion that tasks are subserved by distinct *cognitive* systems (Sherry & Schacter, 1987).

The underlying rhetoric of such approaches is that the allocation of function to separable systems represents an advance in our understanding of particular phenomena (e.g., Ashby, Paul, & Maddox, in press; Evans, 2008; Poldrack & Foerde, 2008). Thus describing a cognitive task as being solved by “System 1” which recruits “region X” of the brain is taken to be a better functional explanation than one which does not include such localisation or identification with a particular system (but see Coltheart 2006; Page, 2006).

Our aim in this article is to examine critically the evidence for and the assumptions underlying such multiple-systems views. We are not the first to undertake such a review, but despite the efforts of those who have pointed out the limitations and inconsistencies of the multiple-systems view (e.g., Gigerenzer & Reiger, 1996; Keren & Schul, 2009; Nosofsky & Zaki, 1998; Mitchell et al., 2009; Palmeri & Flannery, 2002; Shanks & St. John, 1994; Speekenbrink, Channon & Shanks, 2008), it persists as the dominant view and is increasingly presented in popular science as almost accepted ‘fact’ (e.g., Lehrer, 2009).

Our vehicle for the exploration of these issues is human category learning. In recent years increasingly sophisticated multiple-systems interpretations of category learning have appeared (e.g., Minda & Miles, 2010; Poldrack & Foerde, 2008) making category learning an ideal candidate for evaluation. First we present what we interpret as the proponents' 'best' evidence for the existence of multiple category learning systems. We follow each section with a review of studies that challenge the multiple-systems interpretation of some of the empirical phenomena.

In the second half of the article, we take a step-back from the 'for-and-against' interpretations of particular experiments and examine, thoroughly and critically, the assumptions underlying multiple-system accounts. These include assumptions about what we can infer from behavioral dissociations, assumptions about the 'bridges' between neuroscience and behavioral measures, and assumptions about what mathematical models can contribute to our understanding of category learning. We end the article by considering the *varieties of explanation* that can be offered for psychological phenomena and discuss why multiple-systems account can give rise to, what we consider, a false sense of progress and productivity.

### **1.1 Multiple systems of category learning**

Many recent reviews of the literature on human category learning conclude that, "category learning is mediated by multiple, qualitatively distinct systems" (Ashby & Maddox, 2005, p.149), or that the multiple-system approach "is superior in its ability to account for a broad range of data from psychology and neuroscience" (Poldrack & Foerde, 2008, p.197). The over-arching theme of these accounts is that there are two independent systems involved in category learning.

One system, variously termed explicit, declarative, verbal or rule-based relies on working memory, hypothesis testing and the application of simple rules (Ashby & Maddox, 2005; Minda & Miles, 2010). The other, described as implicit, procedural, non-verbal or similarity-based does not involve working memory or attention and learns associations between (motor) responses and category labels (Ashby & Maddox, 2005; Minda & Miles, 2010). The product of learning from the latter system is often assumed to be unavailable to awareness or impossible to verbalise (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Knowlton, Squire, & Gluck, 1994; Minda & Miles, 2010).

Versions of these dual-system accounts abound, and a variety of different tasks have been used in an effort to identify the roles played by ‘explicit’ and ‘implicit’ systems. We focus on two classes of category learning tasks -- probabilistic and deterministic – that differ (as the terms suggest) in the relation between the attributes (cues) of a perceptual stimulus and category membership. From the probabilistic domain we review studies of the popular ‘weather prediction task’ (Knowlton, Mangels, & Squire, 1996), and from the deterministic domain we examine the ‘rule-based’ and ‘information-integration’ tasks popularised by Ashby, Maddox and colleagues (e.g., Maddox & Ashby, 2004). We have omitted discussion of other tasks such as artificial grammar learning and dot-pattern classification, but many of the arguments of interpretation that we raise apply equally to these tasks (see, for example, Nosofsky & Zaki, 1998; Palmeri & Flannery, 2002; Shanks & St. John, 1994).

## **2. REVIEW AND CRITIQUE OF THE EVIDENCE I: PROBABILISTIC CATEGORY LEARNING**

Probabilistic category learning (PCL) involves the presentation of multi-featured stimuli that participants must learn to classify into one of two categories on the basis of trial-by-trial feedback. The feedback is probabilistic thus discouraging participants from memorizing the

outcome associated with a previous encounter with a stimulus. To achieve optimal performance, a participant must integrate information over many trials to establish the appropriate stimulus-response associations. The most commonly used version of the task has a cover story in which participants learn to predict the weather (rainy or fine) on the basis of distinct geometric patterns presented on four individual cards (e.g., a circle card, a square card, etc.). These four cards are presented in all possible combinations (excluding the pattern in which all cards are present on a single trial). Figure 1 shows the basic task along with a brief description of its properties.

[INSERT FIGURE 1 HERE]

The PCL task was developed by Knowlton and colleagues (Knowlton et al., 1996; Knowlton Squire, & Gluck, 1994) to provide a human analogue of the gradually acquired, habit learning tasks used in non-human animal studies (e.g., maze learning; Poldrack & Packard, 2003). The claim that individuals learn the PCL task “without being aware of the information they have acquired” (Knowlton et al., 1996, p. 1400 – see also Gluck, Shohamy & Myers, 2002) fits neatly with the idea that the task is learned by a ‘primitive’ system common to human and non-human animals. Studies of the latter indicate that such habit learning tasks rely strongly on the proper function of the dorsal-striatum (Poldrack & Packard, 2003), thus, in the first studies that used the PCL task it was hypothesized that this area would be important for learning the task.

Since then a good deal of evidence has been collected and interpreted as providing support for this hypothesis. The evidence comes from three principal domains: neuropsychological dissociations, behavioral dissociations and neuroimaging.

## 2.1 Neuropsychological dissociations

A standard practice in cognitive neuropsychology is to identify tasks that can be learned by some patient groups but not by others and then to use these ‘dissociations’ as evidence for the existence of functionally independent systems. The ‘holy grail’ of such investigations is to find a double-dissociation in which patient A can do task 1 but not task 2 and patient B can do task 2 but not 1. Knowlton et al. (1996) claimed to have found a double-dissociation in comparing the performance of amnesics and Parkinson’s disease (PD) sufferers on the PCL task. Amnesics showed unimpaired learning of the task, relative to matched controls, but a clear deficit on a questionnaire that assessed declarative memory for features of the task. In direct contrast, PD patients were impaired at learning the task but showed no deficit on the questionnaire.

Amnesics have damaged medial temporal lobe (MTL) structures whereas PD sufferers have damage to the dorsal-striatum. Thus the double-dissociation reported by Knowlton et al. (1996) supports the notion that these brain structures comprise functionally distinct declarative and procedural memory systems, respectively.

This conclusion was supported in another study with PD patients that demonstrated selective impairment contingent on the type of PCL task used. Shohamy, Myers, Grossman, Sage and Gluck (2004) gave PD patients either the standard version of the PCL task in which learning occurs via trial-by-trial feedback (see Figure 1) or a version in which cues and outcomes are presented simultaneously, and no response is required (an observational or paired-associate version). The PD patients showed the familiar deficit on the feedback version but were relatively unimpaired, relative to controls, in a test following the observational version. This dissociation is consistent with the notion that the dorsal-striatum region is necessary for

feedback learning (and thus PD patients are impaired) but that observational or paired-associate learning relies on the MTL (which remains intact in PD patients).

Shohamy et al., (2004) argued that the selective impairment of PD patients on tasks that rely on learning from feedback is predicted by the underlying neurobiology of the basal ganglia and dopamine system. Specifically, dopamine is released as a result of a prediction-error signal which in turn drives reinforcement learning. When this system is disrupted (as it is in PD) sufferers are unable to learn from feedback.

Taken together, proponents argue that these and similar data (e.g., Reber, Knowlton, & Squire, 1996) provide good evidence for deficits in PCL for patients with dorso-striatal damage but relatively intact performance in amnesics (Poldrack & Foerde, 2008).

## **2.2 Re-evaluating the neuropsychological evidence**

One of the key pieces of evidence in the PCL literature is the claimed double dissociation between PD patients and amnesics reported by Knowlton et al. (1996). Given the elevated status of this study (e.g., Knowlton, 1999; Poldrack & Foerde, 2008; Poldrack & Packard, 2003) it is important to scrutinise whether the conclusions are warranted by the data. On first inspection, the fact that amnesic patients learned the task but were impaired in the declarative memory test whereas PD patients showed the opposite pattern appears to provide good evidence for the involvement of distinct independent systems. But, a closer inspection of the data suggests a more complex situation.

Although the dissociation pattern held for the early part of training (first 50 trials), Knowlton et al. reported that during an additional training period PD patients gradually improved on the PCL task to a level of performance that was the same as the amnesics. However, both patient groups displayed lower average accuracy than controls over the last 50 trials (PD, 61.9%,

amnesics, 59.2%, controls 66.1%). This pattern of data is troubling for the multiple-systems view for two reasons.

First, as Palmeri and Flanery (2002) point out, if the PCL task is a 'process-pure' procedural task then amnesics should be unimpaired relative to controls throughout the task. Knowlton et al.'s explanation for the persistent deficit was that by the end of training PD patients and controls may have been able to access information from declarative memory (i.e., explicit knowledge of cue-outcome associations) that remained unavailable to amnesics. Such an account becomes very difficult to falsify, however, because any unexpected deficits can simply be attributed to the contributions (or lack thereof) from an alternative memory system (Palmeri & Flanery, 2002).

An account in which performance on a task is mediated by multiple systems also undermines a key criterion for the establishment of an independent system – that a system serves a *functionally independent role* (Sherry & Schacter, 1987). If a PCL task can be learnt equally well by either a procedural or a declarative system then it suggests that either the task is a non-diagnostic 'tool' for identifying the correlates of particular systems, or that the system(s), ultimately, is (are) serving the same role (i.e., learning the cue-outcome contingencies in the task environment) (Speenkenbrink et al., 2008).

Speenkenbrink et al., (2008) argued that the difference in levels of performance between amnesics and controls on later trials of the PCL task (e.g., Knowlton et al., 1996) can be captured by differences in the learning rate in a single-system model. Specifically, the learning rate parameter, (which in their associative model determined the size of the change in the direction that minimized the prediction error), was lower for amnesics than controls. Moreover, Speenkenbrink et al. (2008) found in their own study that amnesics and controls had similar high levels of explicit knowledge of cue-outcome contingencies at the end of

training. Thus the learning rate explanation is to be preferred over the suggestion that the later trial advantage for controls over amnesics is due to only the former having access to declarative knowledge (see also Kinder & Shanks, 2003).

The second, related, reason these data are troubling is the demonstration that PD patients *can* learn PCL tasks if given enough time<sup>1</sup>. Wilkinson, Lagnado, Quallo and Jahanashi (2008) drew the same conclusion in a re-examination of the Shohamy et al. (2004) study.

Wilkinson et al. highlighted a number of methodological concerns with the Shohamy et al. design and demonstrated in an experiment that dealt with these concerns that the selective impairment disappeared. Wilkinson et al. suggested that their failure to replicate Shohamy et al. (2004) might have been due to the absence of a response deadline. One of the characteristics of PD is slowness; both of movement and information processing (Ransmayr et al., 1990), and thus the selective impairment Shohamy et al. found on the feedback version of the PCL relative to the observation version may have been due to the timed element and not the feedback component. When the response deadline was removed from both versions, PD patients learned both to equivalent levels. This result undermines the claim that an intact dorsal-striatum is crucial for learning the feedback version of the task. Rather, differences between PD patients and controls reflect variations in response sensitivity/selection not learning per se (cf. Kinder & Shanks, 2003; Nosofsky & Zaki, 1998).

A re-evaluation of some pertinent neuropsychological literature reveals that adopting a one-to-one mapping of deficits in PCL with functionally discrete learning and memory systems in the brain is naïve (cf. Palmeri & Flanery, 2002). Purportedly diagnostic dissociations in which particular patient groups are claimed to be able or unable to acquire different types of knowledge during PCL have been challenged and alternative explanations offered. The most enduring alternative explanation is that all participants, whether brain function is

compromised or not, adopt an explicit rule-based approach to learning the task (Speekenbrink et al., 2008; Speekenbrink, Lagnado, Wilkinson, Jahanshahi, & Shanks, 2010). More evidence for this account can be found from studies examining behavioral dissociations in normal participants.

### **2.3 Behavioral dissociations**

There are surprisingly few studies favoring the multiple-systems view of PCL that rely purely on the presence of behavioral dissociations in humans with unimpaired brain function. An exception to the pattern is a study by Foerde, Poldrack and Knowlton (2007) which examined how PCL was affected in normal participants by the introduction of a secondary memory load task during learning. A straightforward prediction of the multiple-system view is that if PCL is primarily mediated by a procedural learning system then this system should be relatively unaffected by the introduction of a concurrent memory task.

Foerde et al. tested this prediction by giving participants a PCL task to perform on its own or in addition to a secondary tone-counting task. Probe blocks were included at two points in the experiment during which participants who learned under dual-task conditions were released from the secondary task for a given number of trials. Performance was impaired under dual-task conditions, but, when only probe block trials were considered, there was no difference in the accuracy of participants who had learned under single or dual task conditions. Foerde et al. (2007) concluded that the secondary task impacted on the *expression* of learning (i.e., performance) in the PCL task but not on the *learning* of cue-outcome associations.

Moreover, participants who had learned under dual-task conditions had poorer declarative knowledge of cue-outcome associations (in a post-experiment questionnaire) than those who had learned under single task conditions. This pattern of results was taken as supporting the view that declarative knowledge is acquired by the MTL, which is impacted by the additional

secondary load, but that the ability to learn the task is mediated by the dorsal-striatum which is unaffected by load.

## **2.4 Re-considering behavioral dissociations**

There are two reasons to treat the conclusion drawn by Foerde et al. (2007) with caution. First, the design they used compared performance on intermixed 'dual' and 'single' task blocks, but on all blocks corrective feedback was given. Thus it is not clear whether participants learned at a faster-rate in the single-task blocks, or that they were 'free' to demonstrate the knowledge acquired in the previous dual-task phase. Second, Newell, Lagnado and Shanks (2007) demonstrated that dual-task conditions led to a clear impairment of the *learning* of cue-outcome associations in PCL in an experiment in which the number of training trials and no-feedback test trials was matched across load and no load groups. Their data provided no evidence that a release from the dual task during test facilitates the expression of knowledge acquired during training<sup>2</sup>. In a second experiment, Newell et al. (2007) found that participants in a feedback and observation version of the PCL did not differ in their explicit knowledge of the task or the strategies they used to solve it; a result which is unexpected if separate systems mediate performance in the two versions (see also Price, 2009).

A clear prediction of a purely explicit learning account of PCL is that participants will have accurate insight into what they have learnt in the task. Lagnado, Newell, Kahan and Shanks (2006) examined this prediction using an innovative approach in which participants were probed throughout training trials for the explicit basis of each prediction. On each trial participants were asked to rate how much they had relied on each cue in making a prediction. The 'explicit' cue ratings were then compared with the 'implicit' weights derived from

running ‘rolling’ regressions (a series of regressions from predictions to cues across a moving window of consecutive trials).

[INSERT FIGURE 2 HERE]

Figure 2 shows comparison plots of these two dependent measures averaged across participants. In both panels the values are collapsed across the two ‘strong cues’ and the two ‘weak cues’ (see Figure 1 for an explanation of ‘strong’ and ‘weak’). The top panel shows the explicit cue reliance ratings and the bottom panel the weights derived from the rolling regressions. The take-home message from these figures is that participants clearly distinguish between strong and weak cues on *both* the implicit and explicit measures of cue reliance. This ability occurs fairly early in the trials and is maintained, or increases across training. This pattern flies in the face of received wisdom about participants being unaware of what they learn in the PCL task (e.g., Gluck et al., 2002; Knowlton et al., 1996). (An additional experiment demonstrated that overall accuracy in the task was unaffected by the inclusion of the on-line ratings, Lagando et al. 2006, Exp 3.)

The advantage of the on-line measures of cue use is that they interrogate the participant about performance at the time of the prediction rather than at the end of a training episode. Post-training questionnaires that are typically used in studies of PCL often lead to underestimates of explicit knowledge because their retrospective nature leads to distortions of judgment (Lovibond & Shanks, 2002; Shanks & St. John, 1994). Lagnado et al. (2006) also reported strong positive correlations between individuals’ cue reliance ratings and implicit regression weights. The overall pattern strongly suggests that people have access to the internal states that drive their behavior in PCL and that this access drives both on-line predictions and verbal reports. This leaves little room for the contribution of an implicit category learning system.

## 2.5 Neuroimaging

Neuroimaging methods such as functional magnetic resonance imaging (fMRI) have been used increasingly in recent years to examine the neural activity of individuals engaged in PCL tasks (e.g., Aron, et al., 2004; Foerde, Knowlton, & Poldrack, 2006; Poldrack et al., 2001). These studies reveal that the cortico-striatal circuits and mid-brain dopaminergic regions highlighted by the neuropsychological data appear to be actively involved during learning PCL tasks (Poldrack & Foerde, 2008). However, the over-all picture is rather more nuanced.

Several investigations indicate that dorso-striatal regions *and* the MTL are activated when participants are learning, but that the relative involvement of the regions is modulated by a variety of factors. For example, Poldrack et al., 2001 demonstrated that early in learning the MTL was active and the dorso-striatal region (caudate nucleus) was inactive, but as learning progressed this pattern reversed with the caudate becoming active as the MTL deactivated.

In a similar vein, Foerde et al. (2006) showed that when participants learned PCL tasks under single or dual task conditions the striatal learning mechanisms were engaged equally.

However, in a subsequent test-phase in which no feedback was provided, Foerde et al. found that when participants classified items from the PCL task initially learned under dual task conditions, accuracy was correlated with activity in the striatum. In contrast, accuracy for items learned under single-task conditions was correlated with activity in the MTL (right hippocampus).

These results were interpreted as indicating competition between declarative and procedural memory/learning systems. When factors favor the adoption of explicit, declarative learning (such as early in training, or when attention to the task is undivided) then the MTL region dominates. When declarative processes are compromised by additional tasks, or rendered less

important through increased experience with the task, the striatal system takes precedence (Foerde et al., 2006; Poldrack & Foerde, 2008).

## **2.6 Re-imagining neuroimaging**

Neuroimaging studies challenging the multiple-systems interpretation of PCL are rare, if not non-existent. In large part this reflects the interpretative role played by neuroimaging data in the category learning debate. Sections 5 and 6 discuss this role in more depth; here we note simply that studies showing that the relative involvement of MTL and the striatum can be modulated by task demands (e.g., Foerde et al., 2006), do not by themselves constitute incontrovertible evidence for the operation of dissociable systems (cf. Coltheart, 2006; Page, 2006; Sherry & Schacter, 1987). The preceding sections raise several concerns about interpretations of PCL performance and the neuroimaging data should be viewed with these concerns in mind.

## **2.7 Section Summary**

PCL tasks and in particular the weather prediction task have been used extensively in recent years to advance a multiple-systems interpretation of category learning. In spite of apparently compelling evidence, this interpretation can be challenged in each of the areas reviewed. In the neuropsychological domain apparent dissociations between different patient groups dissolve once learning is examined across all trials rather than a subset. Behavioral experiments find that participants have clear insight into their performance, and that the effect of increasing cognitive load is consistent with a single rather than multiple system account. Finally, neuroimaging data while suggestive is by no means conclusive. We now turn to deterministic tasks and argue that similar concerns can be raised there.

### 3. REVIEW AND CRITIQUE OF THE EVIDENCE II: DETERMINISTIC CATEGORY LEARNING

In deterministic category learning participants learn to assign novel stimuli to discrete categories, but unlike probabilistic tasks, feedback is deterministic (a category ‘A’ stimulus is *always* in category A). As noted in Section 1.2, our focus is on two classes of deterministic tasks that have been examined extensively in relation to the COVIS (COmpetition between Verbal and Implicit Systems; Ashby, et al. 1998) model of category learning: rule based (RB) and information integration (II).

In the simplest case, a set of multidimensional stimuli conform to an RB structure if they can be classified on the basis of a single, easily verbalized rule, such as, “If the value of X (e.g., height) is greater than c,” or, “If the value of X is greater than c1 and the value of Y is less than c2”. A categorization problem is II if no easily-verbalizable rule allows perfect classification. A familiar example would be family-resemblance categories, such as those used by Reed (1972), or formed by the faces of members of the Hatfield and McCoy clans. A set of canonical RB/II category structures is shown schematically in Figure 3. Here, the filled squares and unfilled circles correspond to stimuli from two different categories. In Figure 3A, the categories are defined by the level of one relevant dimension (spatial frequency of a perceptual stimulus). In Figure 3B, the categories are defined with respect to the levels of two relevant dimensions (spatial frequency and orientation). An example Gabor Patch stimulus is shown in Figure 3C.

[INSERT FIGURE 3 HERE]

These two classes of tasks are said to engage qualitatively different category learning systems. The RB tasks in which the optimal solution can be described in simple, verbalisable rules recruits the verbal or explicit system which is dependent on working memory and

executive attention (for storing and testing rules respectively). The II tasks for which optimal solutions are difficult or impossible to verbalise (Ashby et al., 1998) does not depend on working memory and attention, but rather learns categories by learning the procedure for generating a response (i.e., the assignment of a category label).

As with PCL tasks, the evidence for this differential involvement of distinct systems comes from three main areas: neuropsychology, behavioral dissociations and neuroimaging.

### **3.1 Neuropsychological dissociations**

Much of the neuropsychological evidence for the involvement of the separable category learning systems identified in COVIS comes from studies of PD patients and so we restrict our review to those studies (see Filoteo & Maddox, 2007; Price, Filoteo, & Maddox, 2009 for more comprehensive reviews). With regard to RB tasks, the degree of impairment shown by PD patients appears to be dependent on the number of *irrelevant* dimensions present in the to-be-categorised stimuli (Price, Filoteo, & Maddox, 2009). For example, Filoteo, Maddox, Ing, Zizak, and Song (2005) gave PD patients and age-matched controls a task in which four binary dimensions of a stimulus could vary trial-to-trial and then manipulated how many of the dimensions were irrelevant for classification. When none or only one of the dimensions was irrelevant, PD patients and controls performed at similar levels. However, when 2 or 3 dimensions were irrelevant PD patients were impaired relative to controls. Filoteo, Maddox, Ing, et al. (2005) concluded that PD patients have deficits in selective attention leading them to be unable to ignore the irrelevant stimulus attributes (see also Ashby, Noble, Filoteo, Waldron, & Ell, 2003; Channon, Jones, & Stephenson, 1993).

In II tasks, in which typically more of the presented stimulus dimensions are *relevant* to the task, one might expect PD patients to show less of a deficit. The data seem to partially support this prediction. Ashby et al. (2003) demonstrated that PD patients performed at a

similar level to healthy controls in II tasks that involved the integration of 3 out of 4 binary valued dimensions. Filoteo, Maddox, Salmon, and Song (2005) contrasted II categories in which the decision bound was linear with those in which it was non-linear. Figure 3B gives an example of a linearly-separable II task; a non linearly separable task is one in which the optimal boundary between the categories is described by an alternative function (quadratic in the case of the Filoteo, Maddox, Salmon, et al. 2005 study). Interestingly, PD patients showed no deficit in learning the linearly-separable II task (replicating Ashby et al, 2003) but were impaired in the more difficult non-linear task. Filoteo and Maddox (2007) speculate that the deficit is specific to non-linear tasks because learning non-linear bounds requires “a greater degree of representation” (p. 16) than learning linear bounds, and this “place[s] more demands on the striatum” (p.16) – an area known to be damaged in PD patients.

### **3.2 Re-evaluating the neuropsychological evidence**

In line with our re-evaluation of the neuropsychological evidence relating to PCL, there are several reasons to be cautious about the one-to-one mapping between deficits and the operation of discrete systems when it comes to deterministic category learning.

Price et al. (2009) note that drawing strong conclusions about the ability of PD patients on RB and II tasks is severely complicated because of the role played by medication (see also Footnote 1). Price et al. (2009) break down the problems faced by PD patients into four issues: *rule generation*, *rule maintenance*, *rule shifting*, and *rule selection*. They conclude that the neurochemical changes associated with PD may disrupt rule shifting (ceasing with an unsuccessful rule after negative feedback) and rule selection (finding an appropriate new rule), but that typical treatment to remedy these deficits (e.g., L-Dopa) can cause detriments in rule generation and rule maintenance. Thus the interpretation of deficits in category

learning depends crucially on whether PD patients are tested ‘on’ or ‘off’ medication (cf. Filoteo & Maddox, 2007)

The PD patients in the Filoteo, Maddox, Ing, et al. (2005) study of RB task learning were all on dopaminergic medication at the time of testing. Thus the impairments seen on the tasks in which 2 or 3 dimensions were irrelevant could have been due to medication causing increased distractability in the patients thereby reducing their ability to ignore irrelevant information and maintain a correct rule (Price et al., 2009). Such a re-interpretation is important to consider because it suggests that it is not the aetiology of PD that leads to impairments, but the nature of the treatment.

A similarly complex picture emerges when one re-considers the evidence for performance of PD patients in II tasks. Price (2005) reported that PD patients on medication were impaired on a linearly-separable II task, contrary to the findings of Filoteo, Maddox, Song et al. (2005) who only reported impairments on non-linearly separable tasks. However, the tasks used in the two studies were very different. The task in Price (2005) involved learning the optimal combination of 5 discrete (presence/absence) cues, one of which was irrelevant to classification. In contrast, Filoteo, Maddox, Song et al. (2005) used a task in which classification was based on correctly integrating the orientation and length of presented lines – thus both dimensions were always relevant. The discrepant results could, therefore, be due to medication impacting on the ability to ignore the irrelevant stimulus dimension in the Price study – something that was not necessary in the study of Filoteo, Maddox, Song et al. (2005).

Other studies showing inconsistent patterns of results also urge caution. For example, Schmitt-Eliassen, Ferstl, Wiesner, Deuschl, and Witt, (2007) failed to replicate Filoteo, Maddox, Song et al’s (2005) finding of PD patient impairment in learning non-linear II tasks

(though differences in experimental procedures may have led to the discrepancy – see Filoteo & Maddox, 2007).

Finally, Swainson et al. (2006) provide some highly diagnostic evidence from a study that contrasted performance of PD patients who were unmedicated or on mild medication.

Participants completed two tasks – the ‘eight-pair task’ – a concurrent discrimination task which involved learning from feedback but not selective attention to dimensions of compound stimuli, and the ‘5-dimensions’ task which required compound discrimination and incorporated trial-by-trial feedback. PD patients – regardless of their medication regime – were unimpaired relative to controls on the eight-pair task; a result which Swainson et al interpret as militating against the claim that the striatum is crucial for feedback learning per se. However, in the 5 dimensions task, mild medicated *but not unmedicated* PD patients were impaired at learning to identify relevant aspects of a compound stimulus. A third group of severely medicated PD patients also performed very poorly on the 5-dimensions task. These results converge with those reported above: PD patients on *medication* are impaired in identifying relevant aspects of multidimensional (or compound) stimuli; however unmedicated patients can learn such tasks (see also Footnote 1).

Taken together, these reconsiderations of the neuropsychological literature suggest that drawing simplistic inferences from the aetiology of PD (e.g., degeneration in dopamine containing cells) to observed deficits on category learning tasks is naïve for two main reasons. First, differences across tasks that are often grouped together as generic RB and II tasks may show very different patterns of performance (Price, 2005). Second, there is clear evidence that PD patients on and off medication behave very differently and in ways that are often at odds with proposed dissociations (Jahanashi et al., 2010; Swainson et al., 2006).

### 3.3 Behavioral dissociations

A huge number of studies document functional dissociations between RB and II categorization tasks in normal, healthy participants (see Ashby & Maddox, 2005; Maddox & Ashby, 2004 for reviews). In a typical experiment, a variable is found to affect learning of either the RB or II structure but to have little or no effect on learning the alternative structure. We focus on four illustrative experiments: two showing the impact of variables on RB (but not II) task learning and two showing the opposite pattern. We acknowledge that this review is selective; however we have chosen those studies highlighted by proponents as providing particular support for the COVIS model (e.g., Ashby et al., in press).

The exclusive reliance of the explicit system on working memory and executive attention led Waldron and Ashby (2001) to predict that increasing cognitive load would have a detrimental effect on RB learning but not II learning. To test this prediction, 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 3 out of 4 of these dimensions were relevant to the classification rule and in the RB task only one dimension was relevant. Participants in the load conditions performed a numerical Stroop task concurrently with the category task. Participants performing the RB and the load task concurrently took longer (more trials) to reach criterion than those performing the RB task alone but load had a negligible effect on II learning. Zeithamova and Maddox (2006) replicated this effect using the Gabor patch stimuli shown in Figure 3.

More support for the claim that RB task learning selectively involves working memory comes from a study by Maddox, Ashby, Ing and Pickering (2004). Maddox et al. manipulated the amount of time participants had to process the corrective feedback delivered after making a categorization decision. Participants were given either RB or II tasks interpolated with a

memory scanning task - identifying a probed digit from a set of 4 briefly presented numbers. In an 'immediate' condition the memory scanning task was presented only 500msec after category-corrective feedback had been given; in the delay condition there was a 2500msec delay. Maddox et al. hypothesized that learning of RB tasks would be detrimentally affected in the immediate condition because there would be no time to explicitly process and reflect on the category feedback. II tasks, in contrast, because they are learned by an implicit procedural system would be unaffected by the availability of feedback processing time. The results provided support for this hypothesis: the RB immediate group performed less accurately than RB delayed, but the II groups did not differ.

The implicit system in COVIS learns categories by learning the procedure for generating a response (Ashby et al, 1998) thus manipulations which interfere with procedural learning should affect performance on II tasks but have little or no impact on RB tasks. Ashby, Ell and Waldron (2003) conducted experiments in which forms of response-motor interference were introduced. After a period of initial learning of both RB and II tasks, Ashby et al. (2003) switched the response buttons used for category assignment. That is, the button that had previously indicated a category A response now indicated a category B response and vice-versa. Following the switch, accuracy in II tasks was affected more than accuracy in the RB tasks – a finding consistent with the predictions of COVIS.

COVIS also makes clear predictions about performance on II tasks based on the underlying neurobiology of the implicit system (Ashby et al., 1998; Ashby et al. in press provide a formal model of this basis). Specifically, the model assumes that II learning is mediated via reinforcement at cortical-striatal synapses, with dopamine serving as the reinforcement signal. For learning to occur, the appropriate synapses need to be strengthened following a reward. This necessitates that some trace of recently active synapses is maintained in the system, but, because of the morphology of the dendritic spiny cells in the caudate nucleus,

this maintenance of activity only lasts for a few seconds. Thus if feedback is delayed it will have an adverse effect on II learning but not RB learning because the explicit system can maintain rules in working memory during the delay.

This prediction was supported in two studies (Maddox, Ashby & Bohil, 2003; Maddox & Ing, 2005) which demonstrated that delays of more than 2.5 seconds had adverse effects on II learning but no effect on learning RB tasks – even when the tasks were matched for difficulty (i.e., number of dimensions relevant to categorization).

These demonstrations, and many others, are taken by proponents of the multiple-systems view as conclusive evidence in favor of discrete separable systems underlying category learning. There is even some evidence for ‘double dissociations’ whereby a single manipulation simultaneously enhances performance on RB tasks and impairs performance on II tasks (Maddox, Love, Glass, & Filoteo, 2008).

### **3.4 Re-considering behavioral dissociations**

As with the neuropsychological data, there are several reasons to question the empirical evidence for separable systems underlying II and RB learning.

Taking the effects on RB learning first, the claim that RB tasks are selectively affected by the addition of a cognitive load is controversial. In an illuminating discussion of the Waldron and Ashby (2001) data, Nosofsky and Kruschke (2002) demonstrated that ALCOVE (a ‘single’-system model) could naturally predict the behavioral pattern observed by Waldron and Ashby (2001) by suggesting that the cognitive load impaired participants’ ability to attend selectively to relevant stimulus dimensions (see Section 5.4 for further discussion of this interpretation). More recently, further doubt has been cast on the interpretation of Waldron and Ashby (2001) via a demonstration that the particular learning criterion they adopted (8

trials consecutively correct) is an unreliable measure of II learning for the stimuli that they used (Tharp & Pickering, 2009).

Zeithamova and Maddox' (2006) replication of the selective effect of load on RB tasks has also been challenged. Newell, Dunn and Kalish (2010) failed to replicate the selective effect in three experiments using a variety of concurrent tasks. 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 concurrent task adequately). Once these participants were removed, all evidence for a dissociative effect of load on RB and II tasks disappeared (see Section 4.4. for more discussion of this study).

In a similar vein, Stanton and Nosofsky (2007) provide an alternative explanation of Maddox et al's (2004) demonstration of the selective effect of a reduction in feedback processing time on RB tasks. Stanton and Nosofsky tested the hypothesis that the dissociation was due to lowered perceptual discriminability of stimuli in the RB structure relative to the II structure. (Perceptual discriminability refers to the distance of items from the category boundary – see Figure 3.) In two experiments they demonstrated that the interpolated memory scanning task had no effect on the RB task when the RB stimuli were easy-to-discriminate, and that II learning *was* affected by the memory scanning task when hard-to-discriminate II stimuli were used. This reversal of the 'dissociation' is clearly inconsistent with the prediction that the tasks are learned by separate cognitive systems.

Nosofsky, Stanton and Zaki (2005) offer a similar 'category complexity' reinterpretation of the Ashby et al. (2003) demonstration that only II tasks are affected by a button-switching manipulation. Nosofsky et al. (2005) argued that because the II category structures used by Ashby et al. (2003) were more difficult to learn than the RB structures, the selective

interference observed for the II task might simply reflect the fact that more difficult tasks are more susceptible to interference than simpler tasks. By manipulating the difficulty of tasks independently of their RB and II status, Nosofsky et al. demonstrated that complex RB tasks were affected *more* than simple II tasks by the button-switching manipulation. Clearly, such a dissociation is inconsistent with the predictions of COVIS and lends support to the idea that task difficulty rather than the recruitment of different systems is the primary mediator of the observed effects.

Such a reinterpretation of the selective effect of delayed feedback on II tasks (e.g. Maddox & Ing, 2005) awaits, but there are grounds to suspect that similar explanations might suffice. For example, in the Maddox and Ing study although the RB and II tasks were matched for difficulty in the sense that the same number of dimensions was relevant for categorization, the II stimuli were less perceptually discriminable than the RB stimuli. When a delay is imposed between prediction and feedback, the feedback is likely to be combined with an impoverished visual representation of the stimulus (cf. Stanton & Nosofsky, 2007). This process will be severely impaired in cases where the initial difference between categories is low (i.e., II stimuli in the Maddox and Ing experiment). Moreover, Maddox and Ing filled the delay between stimulus presentation and feedback with *another* Gabor patch stimulus, thus creating obvious potential for confusing the actual stimulus to which the feedback should be attributed with the ‘mask’ presented during the delay period.

These speculations await further experimental examination but the increasing number of studies reporting findings at odds with the multiple-systems view implies that a more general re-evaluation of the entire hypothesis (of the kind we offer here) is long over-due.

### **3.5 Neuroimaging**

Studies examining the neural correlates of the specific RB and II tasks of the kind shown in Figure 3 are relatively scarce. A relevant study was conducted by Nomura et al. (2007; see also Nomura & Reber, 2008) who scanned participants (using fMRI) while they learnt to categorize stimuli generated from either II or RB structures. The two groups performed identically in terms of accuracy, but Nomura et al. identified some differences in the patterns of brain activation. Successful categorization in the RB task was associated with activation in the hippocampus, anterior cingulate cortex (ACC) and the medial-frontal gyrus, whereas for II classification, correct answers were associated with activation in the head and tail of the caudate nucleus.

Nomura and Reber (2008) explored these patterns of activation further by combining imaging analysis with computational modelling. They defined optimal RB and II models for the tasks and examined participants' data to find sequences of trials on which clear use of RB and II strategies were evident. On those runs where RB behavior was most clearly shown, the right prefrontal cortex (PFC) showed increased activity compared to the best II runs. In contrast, during clear episodes of II strategy use, the right occipital cortical area appeared more active. Nomura et al (2008) interpreted this differential activation as implying a role for working memory (associated with PFC activation) in RB tasks, and the representation of category knowledge acquired by II in the occipital cortex.

### **3.6 Re-imagining neuroimaging**

Although the patterns of activation found in the Nomura et al. (2007; see also Nomura & Reber 2008) study are broadly consistent with the COVIS model they do present some challenges. For example, the original formulation of COVIS (Ashby et al., 1998) did not include the hippocampus as part of the rule-based learning system, instead focusing on the

ACC, PFC and the head of the caudate nucleus. (Note that this last area was more active in the *II* tasks than the RB tasks in the Nomura et al data). More recent expositions do include reference to the hippocampus (e.g., Ashby et al, in press), but seem to suggest that it comprises *another* system separate to the rule and procedural ones described by COVIS (e.g., Maddox et al., 2008).

Moreover, one of the clear messages from the Nomura et al. (2007) data was the high degree of overlap in activation in II and RB tasks. Even when targeted Region of Interest (ROI) analyses were conducted commonality was observed; but rather than interpreting such activity as the operation of a common system it was taken to reflect “the competition between 2 simultaneously active categorization systems” (p. 39). As noted in the discussion of the PCL tasks, invoking notions like the simultaneous operation of systems makes the multiple-systems view very difficult to falsify (cf. Palmeri & Flannery, 2002).

There also appears to be some confusion about the extent to which imaging results from other tasks can be used to support the multiple-systems hypothesis in general and the COVIS model in particular. Ashby and Ennis (2006) discuss imaging data from PCL tasks (like the weather prediction task) as indicating an MTL-based system (rules) and a caudate-based system (implicit) being recruited at different stages of learning the task. However, earlier in the same paper they urge caution about drawing “strong inferences from data collected with [the weather prediction task] because near optimal performance can be achieved by a variety of different strategies (e.g., information-integration, rule-based, explicit memorization)” (p.11). It is as if the neuroimaging data are somehow privileged in illuminating the involvement of separable systems, even when the behavioral data are entirely equivocal on this issue.

Finally, the combination of neuroimaging and computational modelling adopted by Nomura and Reber (2008) is commendable but important limitations on the kind of modelling undertaken preclude strong conclusions from being drawn. We return to this issue in more detail in Section 5.

### **3.7 Section Summary**

Our review and critique of deterministic tasks makes it clear that the wealth of evidence interpreted as consistent with multiple-systems needs to be treated with caution. Evidence from neuropsychological studies is fraught with difficulties because of the role played by medication, and by variations across tasks and experiments. Key studies that demonstrate ‘signature’ behavioural dissociations (e.g. the effect of increased cognitive load, the effect of procedural interference) have been challenged effectively and alternative explanations proposed. Finally, the to-date limited neuroimaging data highlights a good deal of commonality in activation during tasks that are supposedly supported by anatomically distinct systems.

## **4. RE-EXAMINING SOME FUNDAMENTAL ASSUMPTIONS**

As we discussed in the previous sections, much of the evidence that has been used to support multiple-system interpretations of category learning has been based extensively on dissociations in behavioral, neuropsychological and neuroimaging domains. Our aim in this section is to review the logical status of dissociations and to show that they provide a much weaker form of evidence than has often been supposed. As we have argued elsewhere, while it is possible, in principle, to distinguish single and multiple-system accounts<sup>4</sup>, the appropriate logic does not derive from dissociations but rather from *state-trace analysis* (Bamber, 1979; Newell & Dunn, 2008). The present section is divided into three main parts.

In the first part, we present the logic of state-trace analysis and use it to show that dissociations are neither necessary nor sufficient to reject a single-system (or single-process) account of the data. In the second part, we discuss potential difficulties associated with the interpretation of data that are apparently inconsistent with a single-system account. In the final part, we address conceptual issues concerning the nature of the intervening constructs that determine the observed patterns of data. In particular, we examine the question of whether the data alone are sufficient to assert the existence of separate “systems” or “processes” or other theoretical constructs.

#### **4.1 State-trace analysis**

State-trace analysis was originally proposed by Bamber (1979). It is essentially a mathematical analysis of the consequences of manipulating two or more independent variables on two (or more) dependent variables when these effects are mediated either by one, two, or more intervening latent variables. It can easily be shown that qualitatively different outcomes emerge depending upon the number of such intervening variables. For accessible presentations of this logic, in addition to the original article by Bamber (1979), the reader is directed to papers by Loftus, Dillon and Oberg (2004), Newell and Dunn (2008), and Heathcote, Brown and Prince (in press).

Newell and Dunn (2008) explicitly applied the logic of state-trace analysis to experiments in category learning. These experiments generally contrast two different kinds of dependent variables. One dependent variable is usually interpreted as primarily reflecting a procedural or implicit learning system and is operationalized in terms of performance on an information integration (II) category learning task or a PCL task, like the weather prediction task. The other dependent variable is usually interpreted as primarily reflecting a rule-based or explicit learning system and is operationalized in terms of performance on a rule-based (RB) category

learning task or explicit knowledge of the contingencies of the weather prediction task.

Typically, two or more independent variables are also manipulated; one of these is frequently the number of learning trials, the others consisting of factors proposed to differentially affect one or other learning system. An illustrative study is that conducted by Maddox et al. (2003) who examined the effect of various delays between response and feedback on performance on RB and II tasks (discussed previously in Sections 3.3 and 3.4). According to COVIS, delay is critical to procedural learning which depends upon time-dependent strengthening of synaptic links in the tail of the caudate nucleus. In contrast, delay should have little or no effect on rule-based learning which utilizes the storage capacities of working memory.

[INSERT FIGURE 4 HERE]

The theoretical structure proposed by the COVIS model is shown in Figure 4a. According to this model, number of trials affects learning by both the explicit and procedural systems while delay differentially affects learning in the procedural system. In addition, learning by the explicit system primarily determines performance on RB tasks while learning by the procedural system primarily determines performance on II tasks. An alternative theoretical structure, proposed by a “single-system” account, is shown in Figure 4b. According to this model, both number of trials and delay affect a common learning system which differentially affects performance on RB and II tasks. That is, performance on these tasks is considered to depend in different ways upon the same underlying degree of learning. That is, performance is assumed to increase with learning but the rate of this increase may change over time in different ways for the two tasks. Although these changes may be complex, one simple expression of this idea is that one task may be more difficult than the other and therefore increase at a slower rate. In terms of state-trace analysis, this corresponds to the assumption

that performance on each task is a monotonically increasing function of a single intervening variable.

An important analytical tool for state-trace analysis is the state-trace plot. This is a scatter plot of the covariation of the two dependent variables across the different experimental conditions. The two models shown in Figure 4 have different consequences for the form of the state-trace plot. These forms may be classified as either *one-dimensional* or *two-dimensional* (Loftus et al., 2004; Newell & Dunn, 2008). A state-trace plot is one-dimensional if the data points fall on a single monotonically increasing or decreasing curve, otherwise it is two-dimensional. Importantly, the dimensionality of the state-trace plot cannot exceed the number of intervening latent variables. This means that while the multiple-system model shown in Figure 4a may lead to a two-dimensional state-trace plot, the “single-system” model shown in Figure 4b must always lead to a one-dimensional state-trace plot.

[INSERT FIGURE 5 HERE]

Figure 5 shows examples of two state-trace plots. Figure 5a illustrates a two-dimensional state-trace and Figure 5b illustrates a one-dimensional state-trace. In both plots, each point corresponds to the mean level of performance on each of the RB and II tasks for one experimental condition. As these plots are based on the design used by Maddox et al. (2003), there are eight experimental conditions defined by the combination of trial block (one to four, although these are not separately identified in the Figure) and the nature of feedback (immediate vs. delayed). Figure 5a is two-dimensional because the data points do not fall along a single monotonic curve. This is shown by the fact that there are pairs of points (i.e. experimental conditions) across which performance on both RB and II tasks increase and pairs of points across which performance on one task increases and the other decreases.

Three such points are labelled *a*, *b*, and *c* in Figure 5a. Between the conditions corresponding to points *a* and *b* (as well as to points *a* and *c*), both RB performance and II performance increase – the dependent variables are *positively associated* across these conditions. Between the conditions corresponding to *b* and *c*, RB performance increases while II performance decreases – in this case, the dependent variables are *negatively associated* across these conditions. Similar positive and negative associations are repeated between all the other points in Figure 5a. Dunn and Kirsner (1988) called a combination of positive and negative associations a *reversed association*, and showed that it is inconsistent with a one-dimensional monotonically increasing state-trace plot which requires differences in each dependent variable always to be in the same direction.

In contrast, Figure 5b is one-dimensional because in this case all the data points fall on a monotonically increasing curve. While there are pairs of points across which performance on both tasks increase, there are no pairs of points across which performance on one increases and performance on the other decreases. This pattern is shown by the three points, again labelled *a*, *b*, and *c*. Between points *a* and *b*, RB performance increases but II performance remains the same – it neither increases nor decreases. Similarly, between points *b* and *c*, II performance increases while RB performance is constant. Finally, between points *a* and *c*, both RB and II performance increases. Thus, although there are pairs of points across which RB performance and II performance are positively associated, there are no pairs of points across which they are negatively associated. These data therefore fail to produce a reversed association and hence are not inconsistent with a one-dimensional monotonically increasing state-trace plot.

## 4.2 The inferential limits of dissociations

Newell and Dunn (2008) argued that state-trace analysis supersedes the logic of functional dissociation because dissociations are neither necessary nor sufficient to reject a “single system” model of the type shown in Figure 4b. For this reason, dissociations cannot be used as evidence to support an alternative “multiple-systems” model of the type shown in Figure 4a.

A dissociation is defined as the observation that a factor which affects performance of one task has no effect on a second task. In a state-trace plot, performance on one task is plotted against performance on the other. If there is a dissociation across two conditions then this means that the data points corresponding to these conditions will be aligned either vertically, if the affected dependent variable is represented by the  $y$ -axis, or horizontally, if the affected dependent variable is represented by the  $x$ -axis.

Dissociations are not *necessary* to reject a “single-system” model. This is shown by the state-trace plot in Figure 5a. Although this is two-dimensional and thus inconsistent with a single-system account, it contains no dissociations as there are no pairs of data points aligned either vertically or horizontally.

Dissociations are not *sufficient* to reject a “single-system” model. This is shown by the state-trace plot in Figure 5b. Although this is one-dimensional and thus consistent with a single-system account, it contains several examples of dissociations – pairs of data points aligned either horizontally (e.g., points *a* and *b*) or vertically (e.g., points *b* and *c*). Very often, the conjunction of these two kinds of dissociation is referred to as a *double dissociation* and interpreted as providing the strongest evidence in favour of a multiple-systems account (Shallice, 1988). However, this example shows that a dissociation (even a double dissociation) is not sufficient to reject the “single-system” model.

State-trace analysis supersedes the logic of dissociations because it specifies a pattern of data that is logically inconsistent with a single-system account. In contrast to single and double dissociations, a two-dimensional state-trace plot is both necessary and sufficient to reject a single-system account.

State-trace analysis also implies a different set of statistical questions. A feature of dissociation logic is that it often depends upon showing that conditions do *not* differ on one or more dependent variable thereby running the risk that any conclusions that may be drawn may be a consequence of one or more Type I errors. In contrast, in order to reject the “single-system” model, state-trace analysis requires that pairs of conditions should differ in systematic ways. Specifically, it is necessary to show that there is at least one pair of points that differ in the same direction on *both* dependent variables and another pair of points that differ in opposite directions on both dependent variables (Dunn & Kirsner, 1988; Newell & Dunn, 2008).

[INSERT FIGURE 6 HERE]

The statistical analysis of state-trace plots is currently under active investigation (Heathcote et al, in press; Newell et al., 2010). The point we wish to make here is that different conclusions may be drawn from the same data depending upon whether it is analysed in terms of the logic of dissociation or of state-trace analysis. Figure 6 shows the state-trace plot of data from Experiment 1 of the study by Maddox et al., (2003) averaged over three levels of time interval<sup>5</sup>. Two features are apparent in these data. First, as concluded by Maddox et al, and noted above, these data demonstrate a double dissociation – there are pairs of data points aligned, at least approximately, both vertically and horizontally. Second, the data appear to be consistent with a one-dimensional state-trace plot. In fact, the data points shown in Figure 5b correspond to the best-fitting monotonic curve passing through the observed data shown in

Figure 6 (for details of how to fit such a curve, see Newell et al, 2010). Although a firm conclusion depends on formal statistical analysis, it is highly unlikely that the small differences between the observed data in Figure 6 and the best-fitting monotonically increasing data in Figure 5b are sufficient to reject a “single-system” account. These data therefore offer little or no support for a multiple-systems view.

### **4.3 A state-trace re-analysis of behavioral and other dissociations**

We have proposed that arguments for or against multiple-system accounts of category learning should move beyond dissociation logic. As illustrated above, re-analysis of at least one prominent behavioral dissociation using state-trace analysis shows that a quite different conclusion may be drawn from it. This does not imply that all or any other dissociation is open to a similar re-interpretation. While it is beyond the scope of the present article to examine each dissociation in turn, it is at least possible that some may also be shown not to meet the higher test offered by state-trace analysis. The view that the case for COVIS or other multiple-systems accounts of category learning has been well established is therefore based on an interpretation of the evidence that is currently only provisional.

The arguments we have proposed above apply with equal force to dissociations in other domains. For example, in a review of the role of the basal ganglia in category learning, Ashby and Ennis (2006) have drawn attention to dissociations in category learning performance between different patient groups. They reviewed evidence that patients with Huntington’s disease, a degenerative disease that affects most of the basal ganglia, are impaired on both RB and II tasks compared to normal controls. In contrast, patients with Parkinson’s disease, a degenerative disease that primarily affects the head of the caudate nucleus, are impaired on RB tasks but are relatively less impaired on II tasks (see Sections 3.1 and 3.2 for further discussion). This dissociation is consistent with the COVIS model that

proposes that the explicit system that underlies learning in RB tasks is subserved by the head of the caudate nucleus while the implicit system that underlies learning in II tasks is subserved by the tail of the caudate nucleus.

Since both patient groups are impaired on RB tasks, evidence for the critical dissociation depends on differences in performance on II tasks. In one study, Filoteo, Maddox & Davis (2001) compared Huntington's disease patients with normal controls on two different II tasks that varied in difficulty. In the easier task, the categories were linearly separable, in the harder task, they were separated by a more complex nonlinear boundary. The patient groups were impaired in learning both kinds of task. In a later study, Filoteo, Maddox, Salmon et al. (2005) compared Parkinson's disease patients with normal controls on two, similarly defined II tasks. In this case, the Parkinson's patients were impaired on only the more difficult task (as discussed in Sections 3.1 and 3.2).

[INSERT FIGURE 7 HERE]

As noted earlier, dissociations depend upon the failure to reject the null hypothesis of no difference and are thus always subject to a Type I error. For this and other reasons, we have argued that the appropriate statistical approach should be based on state-trace analysis. Figure 7 presents a state-trace plot of the results reported by Filoteo et al (2001) and Filoteo, Maddox, Salmon, et al (2005)<sup>6</sup>. It shows performance on both the linear (easy) and nonlinear (hard) II tasks for each of six blocks of trials for the two sets of patient groups and their corresponding normal controls. Even though the stimulus sets and the nature of the nonlinear bound differed between the two experiments, the data nevertheless appear to fall on essentially the same monotonically increasing curve. While this may be a coincidence, the most important point is that this plot reveals that the differences in performance between the

two patient groups can be accounted for by differences in the relative difficulty of the II tasks between the two experiments and the shape of the resulting curve.

The shape of the curve may be interpreted as resulting from different changes in the performance on the two kinds of task as a function of learning. Because these changes are nonlinear, apparent dissociations may appear. In the case of the Parkinson's patients, the difference in their performance relative to the controls appears to be greater on the nonlinear II task than on the linear II task. In the case of the Huntington's patients, the difference in their performance relative to the controls appears to be approximately the same on both tasks (although the shape of the curve also suggests that they may be relatively more impaired on the linear task). Therefore, while these data may be informative in other ways, state-trace analysis shows that they do not offer strong evidence against a "single-system" model.

#### **4.4 Interpretation of two-dimensional state-trace plots: The role of confounds**

The "single-system" model shown in Figure 4b is a convenient *straw man* that is, in our opinion, unlikely to be true. This does not mean that we endorse an alternative multiple-systems view. Rather, we are mindful of the fact that any two tasks or dependent variables that are sufficiently dissimilar to warrant investigation are likely to draw upon a (potentially large) number of different systems or processes or be affected by other uncontrolled variables. For this reason, showing that two tasks can, under some circumstances, yield a two-dimensional state-trace is an essentially trivial result since it is necessarily true. The challenge in this respect is to show that the two tasks yield a two-dimensional state-trace under conditions that are theoretically relevant. This is not always obvious nor an easy thing to achieve.

One important advantage of a model such as COVIS is that it clearly specifies a set of minimal contrasts that should differentially affect the different category learning systems it

proposes and thereby yield a two-dimensional state-trace plot. Nevertheless, the attribution of such a pattern of results to the relevant theoretical constructs is not always straightforward.

This concern is nothing new and is simply the extension of the principles of good experimental design to the bivariate domain of state-trace analysis. For this reason, just as in the more familiar univariate domain, it is necessary to mount an additional argument to show that an observed result can be attributed to the theoretical constructs of interest.

The potential pitfalls associated with the interpretation of two-dimensional state-trace plots can be illustrated by a recent study by Newell et al. (2010)<sup>7</sup>. The starting point for this study was the observed dissociation, reported by Zeithamova and Maddox (2006) that category learning in an RB task is selectively impaired by concurrent task demands compared to learning in an II task. Re-analysis of these data using state-trace analysis revealed an apparent two-dimensional state-trace plot. Although it was not possible to reject a one-dimensional curve on the basis of formal statistical analysis, the original study was not designed with this analysis in mind and there was enough evidence, in our minds, to suggest the involvement of multiple systems consistent with the predictions of COVIS. However, in a subsequent series of replications of the basic experiment we consistently observed unambiguous one-dimensional state-trace plots. On examination of this apparent inconsistency, Newell et al. observed that, in their experiment, Zeithamova and Maddox did not partition their participants into those who appeared to learn the task and those who did not. When only the learners were analysed, the data from Zeithamova and Maddox (2006, Experiment 1) also revealed a clear one-dimensional state-trace plot, consistent with the pattern of data found by Newell et al.

The inclusion of non-learners was sufficient to change the dimensionality of the state-trace plot. As pointed out by Newell et al, this was due to the fact that different proportions of non-

learners occurred under different conditions of the experiment. That is, the proportions of non-learners varied between the load and no-load conditions and between those learning either a RB or II structure. This meant that although the performance of learners was well-described as a function of single learning parameter, consistent with a “single-system” model, the performance of the entire group was a function of both the amount learned and a variable proportion of non-learners who, by definition, learned nothing. Because this proportion differed between conditions, the resulting state-trace plot was approximately two-dimensional leading to the incorrect inference that performance was a function of two latent variables or learning systems<sup>8</sup>. This example suggests that caution should be exercised in attributing an observed two-dimensional structure to the theoretical mechanism of interest. As in any experiment, confounds are possible (See Section 3.4 for discussion of other ‘confounds’ in experiments comparing II and RB tasks, e.g., Nosofsky, et al. 2005).

#### **4.5 Interpretation of two-dimensional state-trace plots: Systems, processes and latent variables**

Dissociation logic was initially developed in neuropsychology (Lackner, 2009; Teuber, 1955). In that context, dissociations were naturally interpreted in terms of major functional capacities, such as reading, writing, language, and planning, involving relatively large brain regions and functional processing systems. This manner of interpreting dissociations has continued to the present day and forms one of the methodological foundations of the cognitive neuropsychology program where it is used extensively to partition mental function into separate processing *modules* (Coltheart, 1985; Shallice, 1988). However, there is nothing in the logic of dissociations that compels this interpretation, a point that is made explicit in the logic of state-trace analysis. In this context, the intervening constructs are best described as latent variables or model parameters that have no substantive meaning outside of the

particular theory to which they are relevant. This point of view was first proposed by Dunn and Kirsner (1988) in their analysis of dissociation logic in which they represented the effects of independent variables on latent variables and the effects of latent variables on dependent variables simply in terms of mathematical transformations between one set of variables and another. Although they referred to the intervening variables as “processes”, these were viewed as abstract constructs or parameters that had no essential meaning.

It follows from the above that evidence for multiple intervening constructs, whether based on dissociations or, as we propose, state-trace analysis, is equally consistent with any number of different theoretical interpretations. It is thus equally consistent with an interpretation based on multiple systems, such as COVIS, as with one based on multiple processes or parameters, such as ALCOVE (Kruschke, 1992).

ALCOVE proposes that category learning is dependent on a single psychological system governed by four different control parameters. In any one study, if at least two of these parameters are differentially affected by the independent variables and if they, in turn, differentially affect performance on RB and II tasks, then a two-dimensional state-trace plot (as well as potential dissociations) will result. The dimensionality of the state-trace plot reveals the number of underlying latent variables but says nothing about their nature. In particular, state-trace analysis does not offer a principled means of distinguishing between an interpretation of the data in terms of multiple parameters of a single system (as per ALCOVE) or in terms of parameters of multiple systems (as per COVIS). Distinguishing between these interpretations requires additional criteria such as the nature of experimental effects, their internal logic and their respective abilities to account for the data. This, we suggest, depends upon specification of an explicit mathematical model that precisely defines

the theoretical constructs in a form that allows for quantitative evaluation. It is to this issue that we turn in the next section.

## 5. THE CONTRIBUTION OF MATHEMATICAL MODELLING

In the context of a discussion contrasting the multiple- and single-system view of categorization, formal models would seem to be an important touchstone. For example, Ashby and Ell (2002) offer the following definition of a system: a system is a mapping of state variables that represent inputs, to state variables that represent outputs, in a manner governed by a vector of parameters. The notion of a system, then, is fundamentally formal. For two systems to be different there must be some parameter values that produce mappings that are unique to each system; if this is the case then the systems are neither identical nor fully nested. For two systems to be systems of the same type (and so potentially ‘the same’), the input and output variables must be the same. Ashby and Ell further propose that for any candidate systems to be candidate *behavioral* systems each must make observable responses. This requirement is taken by Ashby and Ell to be met if the candidate systems produce outputs that can be interpreted as response choices.

As suggested in Section 4, any model with multiple parameters can, in principle, produce observable dissociations. The generality of this statement suggests that any initial optimism about mathematical models *per se* resolving a multiple-systems debate should be tempered with a healthy scepticism about the ability of models to, on the one hand, uniquely describe a pattern of selective influence and, on the other, to unambiguously qualify as models of single or multiple systems.

In this section we briefly try to determine whether the COVIS model (as an exemplary multiple-system model) has advanced the multiple-system discussion.

### 5.1 Mathematical/computational models of category learning

Excellent, contemporary, reviews of formal models of category learning are readily available (e.g., Kruschke, 2005, 2008). The primary dimensions that differentiate formal models are the sorts of representations, the sorts of learning rules, and the sorts of response-generation processes they employ. The last of these is the least theoretically relevant to category learning, the second is critical for understanding the trajectory of categorization (do people learn more on trials when they are incorrect, or do they learn equally on all trials, or do they update their beliefs rationally, etc.). It is only the first, the representations a model postulates, that is of central relevance to the question of multiple systems.

The kinds of representational schemes proposed for categorization break down naturally into two kinds. A pure exemplar model would propose that all items presented for classification are stored in memory, and that classification results in consulting all of these items. A pure abstraction model would propose that only a summary of the presented items is retained; this summary could be a theory, a rule, a prototype, or a boundary of some variety. Impure, or hybrid, models mix these representational formats. Examples of hybrid models include COVIS, RULEX (Nosofsky, Palmeri & McKinley, 1994), ATRIUM (Erickson & Kruschke, 1998), SUSTAIN (Love, Medin, & Gureckis, 2004), varieties of Rational models (Anderson, 1991; Tenenbaum & Griffiths, 2001), VAM (Vanpaemel & Storms, 2008), Smith and Minda's (2000) approach, Minda and Miles' (2010) model, and probably more all the time. Some of these have been very successful in accounting for a variety of data, others far less so.

Hybrid models pose a natural group for consideration as multiple-system accounts (leaving aside the question of how one could hope to identify such an account, given the flexibility of any multiple-parameter model). This appearance is deceiving, however, as many hybrid accounts either explicitly deny the possibility, or do not meet the criteria for multiple systems. The first case is presented in Kruschke's (2008) thoughtful review in this way,

“[The] representational assumption for a model does not necessarily imply that the mind makes a formal representation of the stimulus. Only a formal model requires a formal description... The representations in the model help us understand the behavior, but those representations need not be reified in the behavior being modelled.” (p.269, fn1). If the representations are only in the model, and not in the brain, then they cannot determine systems of categorization, because it is only the person, and not a formal model, that can produce behaviors.

The second case is exemplified by a model like SUSTAIN or VAM, where the component representations are unified into a single mechanism, and do not produce distinct formal predictions. It is only COVIS that claims to be a genuine multiple-systems model.

## 5.2 COVIS

The original formulation of COVIS (Ashby, et al. 1998) as a neuropsychological theory was supposed to span all three of Marr’s (1982) levels of description of an information-processing system: task (or computation), algorithm, and implementation. Marr’s notion of a task description is the ideal-observer solution that describes performance; that is, there must be some set of sources of information and some rules for their combination that describe the relationship between the stimulus and the response and these sources and rules constitute the ‘task’. Ashby et al. provide a different sort of task analysis, by characterising the global dynamics of COVIS, with a particular focus on the bias the model predicts due to the way it integrates its rule-based and implicit-boundary-based subsystems<sup>9</sup>.

The implementation level description of COVIS identifies its formal components (inputs, subsystems, system-combination, response generation) one-for-one with brain regions.

COVIS (as presented most recently in Ashby, et al. in press) is a complicated model. It has at least 19 free parameters, at least 8 of which govern the verbal system. It has at least five

random factors, which prevents the model from making deterministic predictions, and the competition between the verbal and procedural systems is a complex relationship depending on the performance of each system and the biases of the model as a whole. What COVIS predicts for a given experiment cannot be determined without investigating its parameter space.

Thus, predictions from COVIS about the effects of various lesions or diseases on categorization are by virtue, first, of the identification of the formal structure of the capacity to categorize and, second, of the correct identification of the parts of the brain that support (vs. “implement”) the various aspects of that structure. This order is critical – if a model does not get the formal structure correct (if it can’t fit data), then its claims about identity with brain regions are necessarily irrelevant (if it can fit data, then its claims may remain senseless!). Thus, our concern here is with the formal structure (what Ashby et al. take as the algorithmic level<sup>10</sup>) of the model, to see how it fits data.

### **5.3 The nature and use of formal models**

While the utility of formal models in science is beyond doubt, their nature and use in psychology is problematic. Lewandowsky and Farrell (in press) remind us of the paradigmatic case from astronomy, in which planetary motion models moved from Ptolemaic to Copernican to Keplerian due to two considerations. The first shift was due primarily to the simplicity of the Copernican model, which did not fit the data much better than its geocentric competitor, while the second shift was due to the near-perfect fit of the Keplerian model to the then-available data. The example is attractive for another reason: as Newton showed, the Keplerian model is consistent with an underlying causal mechanism (gravity). We now understand this model to be true: The solar system moves the way it does because its bodies have the masses, positions, and velocities that they do. In psychological theory, the parameter vectors of formal models take the role of constructs like ‘mass’, but we still wish

to know whether any particular model is closer to the truth than any other. The promise of a neuropsychological model like COVIS is that it seems to be better positioned to be true than a (mere) psychological model, because it specifies something extra, something akin to adding ‘gravity’ to the explanation of planetary motion.

The argument over single- vs. multiple-systems has developed (though it need not have) partially around this distinction. Single-systems psychological models, like the GCM and ALCOVE, do not claim that their formalisms are actually entities in some inner mental world (as the quote earlier by Kruschke, 2008, makes clear). Their truth is established only by behavioural experiments; they are in essence *measurement models* (Stevens, 1946) that tell how a parameter can be measured with an experiment. Neuropsychological models like COVIS, on the other hand, do claim that their formalisms are entities; that the components of the model are regions or states or processes in the brain. Their truth is jointly established by behavioural and neuropsychological experiments. The implications of this for evaluating COVIS are difficult to tease out.

COVIS, just as it claims to be, is two kinds of model at once. On the one hand it is a formal measurement model, with equations specifying how its parameters account for data. These equations tell us, for example, how increasing a learning rate parameter, or a decision rule parameter, will change the model’s predicted responses. On the other hand, the parameters are provided meaning by being embedded in an informal theory, a kind of *structural model* (Busemeyer & Jones, 1983), so that some tasks, or some factors (like WMC, or brain diseases) are taken to cause the parameters to change in certain ways. The presence of a concurrent task is supposed to interfere with the explicit system, reducing some parameters and increasing others, while leaving the implicit system alone. This distinction is clear, in principle, but becomes unclear in practice due to the ambiguous role of the neuropsychological explanations.

The point of raising the issue of measurement vs. structural models is to provide leverage on the question of the role of computational modelling in establishing the veracity of the multiple-systems theory. This leverage is partially undercut by the dual role of neuropsychological theory in the model – it enters at both the structural level (where, e.g., the locus of the explicit system is identified) and the measurement level (for example, where the values, or at least names, of parameters in the implicit system are identified). Thus, while the presence of a formal model might suggest a unity between psychological and neural processes akin to the relationship between the geometry of Kepler and the physics of Newton, in the case of COVIS the relationship is distinctly unclear.

To clarify the relationship between the neural and the psychological aspects of COVIS, let us look at a plausible prediction the model might make: Learning to discriminate on the basis of family resemblance (e.g., to tell the Hatfields from the McCoys) depends on the timely delivery of dopamine to the procedural learning system. This prediction involves three things we don't know – and three ways experimentation might enlighten us about categorization. At the highest level, the claim refers to an empirical and testable statement about the world, “learning family resemblances depends on dopamine”. This requires two measurements: one the degree to which a person has actually learned a family resemblance (vs. e.g. learned a simple rule, or not learned at all), and the other a measure of the availability of dopamine. Should this relationship be established, we would then be left with two more model-theoretic, (potential) explanations of this fact. The first of these refers to a psychological model of the mental operations that underlie learning family resemblances and states that such learning is “procedural”. The second claim refers to a neuropsychological theory (which is also a model but we will use different terms to avoid confusion) about how the set of operations described by the psychological model depend on or are subserved by physical properties of the brain. This theory states that procedural learning (whatever that might be) depends on dopamine.

What is the relationship between the psychological model and the neuropsychological theory? One possibility is that of entailment. For example, our understanding of brain science may be advanced enough to know that if learning is procedural then it has to depend on dopamine, as there is no other way this kind of learning could be implemented in the brain. The neuropsychological theory is thus entailed by the psychological model and is a bona fide prediction of this model. This means that were it to be shown that learning family resemblances does not depend on dopamine, it would follow that it cannot be based on procedural learning.

It is highly unlikely, given our present understanding of brain science, that in COVIS or other similar models, the neuropsychological theory is entailed by the psychological model (Coltheart, 2006; Uttal, 2001). If this were so then the claim that category learning depends upon two separate learning systems would depend upon particular neuropsychological facts. For example, if it were shown tomorrow that parts of the caudate nucleus were not differentially involved in learning rule-based and non-rule-based categories, would this challenge the psychological claim that there are two different learning systems? If it did then what are we to make of the large number of behavioral dissociations used to support this hypothesis? Either they do not distinguish between single-system or multiple-system views, in which case they cannot count as evidence one way or the other, or they do, in which case they do so independently of how they may be implemented in the brain. That is, either the neuropsychological data are irrelevant to the claim or the behavioral data are irrelevant.

Since it is claimed that neuropsychological and behavioral data are both relevant to the evaluation of COVIS, it follows that the relationship between the two parts of the model cannot be one of entailment (as defined above). What else could it be? One possibility is that of contingency. That is, as far as we know, it happens to be an empirical fact that the rule-

based and procedural systems depend on proper function of the head and tail of the caudate nucleus (for example). In this case, were it shown tomorrow that these parts of the caudate nucleus are not differentially involved in learning rule-based and non-rule-based categories, the psychological model would be unassailed with the neuropsychological theory being simply revised accordingly. It follows from this view that while neuropsychological facts may be a useful source of hypotheses concerning the functional characteristics of the psychological model, nothing in the final analysis depends on this – neuropsychological data are ultimately irrelevant to the central claims of the psychological model.

It follows from the above argument that unless one were to want to make the strong and, at our present understanding of brain science, unwarranted claim that the psychological model part of COVIS entails the neuropsychological theory part then the only evidence for the former derives from the behavioral data and its ability to account for these data.

#### **5.4 Model comparisons**

How well does the formal psychological model part of COVIS, sketched above, do at capturing performance, relative to other models? The short answer is, we don't really know. A full description of the formal structure of the model is beyond the scope of this article. Interested readers should consult Ashby et al. (in press) for an in-depth presentation. However, what is clear from the full description of the model is that because it has so many stochastic components (e.g., which rule is chosen for switching, how likely a person is to switch to a new rule, which rule is chosen to be active, what the active rule predicts for a response, what the procedural system predicts for a response) interacting in non-algebraic fashions, it is probably impossible to fit to data using MLE methods. Demonstrations of fit are made via MonteCarlo sampling (Ashby et al. 1998), but are not very frequent.

Ashby et al. (in press) provide another example of this, in their demonstration of COVIS's ability to account for behavioral data by approximating the dissociation observed by Waldron and Ashby (2001) (discussed in Sections 3.3 and 3.4). Ashby et al. say that, "because the dual task reduces the capacity of working memory and executive attention, we assumed that participants would therefore be less adept at selecting appropriate new rules to try and at switching attention away from ineffective rules." Thus, the demonstration of the model's ability to accommodate the data proceeds by modifying two of the explicit system's parameters (one that governs the chance that the system will persevere in the face of error, and one that governs the chance that the system will select a random rule instead of a salient rule). The demonstration also involves, due to the nonconfusable nature of the stimuli, both preventing any generalization in the implicit system and making the rule system's predictions deterministic. This produces the predicted (and observed) near-perfect dissociation, with almost no effect of concurrent task on performance in the II condition. Critically, the model parameters are not discovered by a statistical procedure, but by the application of an informal structural model.

It is informative to compare the Ashby, et al. (in press) model-based account of the Waldron and Ashby (2001) data with Nosofsky and Kruschke's (2002) account. Nosofsky and Kruschke explored the model ALCOVE, which is a four-parameter model combining exemplar representations, dimensional attention, and error-driven learning. Nosofsky and Kruschke found that ALCOVE could produce results like Waldron and Ashby's over a wide range of parameter values, simply by allowing that the concurrent-task interfered with participants' ability to learn to modify their dimensional attention. This would be generally equivalent to COVIS losing the ability to use its verbal rule system, although this is not Ashby et al.'s preferred explanation via COVIS.

It is worth noting that ALCOVE could only fit the data when the generalization was rather broad, allowing associations learned for one stimulus to generalize to others even though the items were not confusable. Ashby et al.'s demonstration was obtained in the exact opposite situation, where the generalization was very narrow. This emphasizes Nosofsky and Kruschke's point that, "[i]t seems advisable to test alternative models by using richer sets of parametric data, rather than relying on qualitative patterns of results involving just two data points. Before ruling out an entire class of models, one should conduct an investigation over a reasonably large region of the models' available parameter space." (p171). COVIS (the formal model) has not been tested against data sets in this manner.

The argument against formal single-system accounts of category learning is generally that they do not predict the dissociations multiple-system theorists explore. This has nothing to do with the formal models of either single or multiple systems, but entirely to do with the informal theories that surround them. Mathematical models can resolve only so much of the debate about multiple systems. Parameter estimation can tell us which parameter values are required to approximate the data from any given set of manipulations, and informal theory (or common sense) can tell us if these values are meaningful. Model selection can tell us whether one model is generally more accurate given its flexibility, but model selection is incredibly difficult given the complexity and similarity of the formal models and the variability inherent in human category learning.

In summary, examination of COVIS as an example of a multiple-systems mathematical model leaves a number of questions unanswered. First, due to the nature of its formalisms it is very difficult to compare its performance against other models, and thus such rigorous comparisons have yet to be undertaken. Second, the dissociations in behavioral data do not provide unambiguous support for COVIS, and some at least can be explained via other 'single-system' models. Third, one of the key selling points of the model –its ability to

account for phenomena at both the neural and the psychological level – make it unnecessarily complex, especially when ultimately the neuropsychological data are irrelevant to the debate about how many systems are involved in our *capacity* to categorise.

## 6. DISCUSSION AND CONCLUSIONS

The title of this article is deliberately provocative and, of course, a little flippant. However, what we hope to have highlighted in the preceding sections is that determining the ‘fact’ or ‘fantasy’ of discrete, functionally independent systems of category learning requires very careful consideration and integration of a wide range of evidence.

This evidence suggests to us that the case for multiple systems underlying either probabilistic or deterministic category learning is still very much open to debate and may be challenged in three main ways. First, in our review of the “best” empirical evidence supporting a multiple-systems interpretation, we observed that the interpretation of much of this research is contentious and open to alternative explanations (Sections 2 and 3). Second, we have argued that since much of this research has relied on the discovery of dissociations, it rests on a flawed logic (Section 4). A feature of this logic is that it depends on the failure to observe an effect and thus is particularly prone to Type I errors. This means that researchers may be easily misled into believing that they have discovered multiple processing systems when they have simply failed to observe a significant effect.

Finally, we have argued that dual-nature of theoretical accounts such as COVIS that link a psychological model of category learning to proposed neuropsychological structures and processes may serve to complicate rather than to clarify our understanding of category learning at both levels (Section 5). At a fundamental level, the fact that data can be fit by a mathematical model that incorporates multiple systems in no way reifies the existence of these systems in the person learning to categorise. Ultimately, however, the answer to the

question we pose in the title requires clarity about what constitutes an explanation of a given phenomenon.

### **6.1 Varieties of Explanation**

Psychological theory is a very difficult project, and is subject to problems other scientific theories easily avoid. Every field must be clear about what it seeks to explain, and how it seeks to explain it. Substituting an explanandum for an explanation is wrong, as when an effect becomes a cause – e.g., when someone identifies ‘the Stroop effect’ as the reason people are slowed by contradictory information. These may pass notice because the effect is a technical term, and technical terms are often explanations. But in psychology we also must guard against ordinary terms, like ‘rule’, being mistaken for technical ones. So when we wish to explain how a person learns a category, and we want to say ‘by testing rules’ we must realize that we have not offered a technical explanation, but an ordinary (perfectly sensible) one. There is only one kind of ‘system’ that can follow a rule, and that is an intelligent creature who can be said to follow it correctly, or not. A volcano cannot erupt in violation of a rule, though its eruption can be inexplicable or unusual. A neuron cannot fire correctly, but only normally. A person can follow a rule, and can forget to do so.

Explanations of things that people do are fraught, because we have two ways of offering such explanations. As Bennett and Hacker (2003) so lucidly describe, consider a man who buys a pair of opera tickets. Why does he do so? The explanation, “Because he wishes to give his wife a present” is perfectly valid. His wishes are constitutive of the act of buying the tickets (as opposed to buying them by accident, or being tricked into it). The explanation, “Because his brain was in such-and-such a configuration” might be valid if we understand that the brain is not constitutive (it doesn’t enter into what we mean by ‘buy the tickets’). Instead, the brain

explanation is a description of the causally enabling conditions (Trigg & Kalish, 2009) for exercising the capacity to e.g., buy the tickets.

In the same way, a formal model can be of one of two types. One can model the constitutive components of a behavior, as we might when modelling how people make choices, reach decisions, test hypotheses, rely on heuristics, or draw conclusions from arguments (e.g., Evans, 2008; Gigerenzer & Todd, 1999; Tversky & Kahneman, 1981). In the case of category learning, such a model could quantify the probability that a person considers some particular simple rule about single-features before a more complicated two-feature rule, for example.

The other variety of explanation is to model the causally enabling conditions of a behavior, as we might when describing category learning as the association of labels with stored instances. All this kind of model amounts to is a mapping of measured inputs to measured responses; an answer to the question raised in Section 5.3 about what the word ‘properly’ in the phrase ‘the brain must be functioning properly’ means. Models of perception, memory retrieval and familiar models of category learning (e.g., ALCOVE) can be read as of this variety, as they explicate the mechanisms that support our capacity to perceive, remember and learn. The possibility that psychology might arrive at this variety of model is thrilling, and to mistake it with the first type, only confusing.

Multiple-systems theories of categorization, such as COVIS, are uniformly of this confused sort since they mix constitutive and causal explanations as if they are both the same kind of explanation. On the one hand we have an explicit ‘system’, which is really a system that does things that a person does. Only a person can use language (so a ‘verbal system’ is a person), or follow a rule (for a ‘rule-based system’), or represent something to him/herself in an explicit way. Explanations based on this ‘system’ can become meaningless: if the explanation of how a person learns a rule is that their ‘verbal system’ learns the rule, then

either the explanandum has simply served as an explanation, or the explanation is homuncular. On the other hand multiple-systems models postulate an implicit ‘system’, which is a description of how our brain has to be organized such that we can learn things like how to shoot a basketball, or go left on a wave, or, maybe, tell a Hatfield from a McCoy. The difference between these kinds of explanations can be seen in what would happen if they were to fail. If the implicit ‘system’ were broken we could not do the things we do, but if the verbal ‘system’ were broken *we* would not be *doing* anything at all (things might still be happening to us or in us...). For a multiple-system theory to make sense (let alone to be true or false) it must respect the difference between who we are, and how we work.

An interesting consequence of this distinction is that respecting it does not eliminate theories that propose qualitatively different ways to learn a categorical distinction. The formal model that is COVIS might even be just such a theory if it were interpreted correctly – that is, if it made no reference to ‘explicit’ vs. ‘implicit’ systems, and if the actions of the rule-based system were understood as abstractions of function. Many of the hybrid models of section 5.1 are compatible with just this sort of interpretation, as indeed are even simple models (like ALCOVE) that depend qualitatively on the values of their parameters. But respecting the distinction eliminates the logical possibility of multiple-systems theories that contrast explicit and implicit ‘systems’.

## **6.2 Explanatory power**

One of the clear impediments to resolving the multiple- vs. single-systems debate in category learning (and many other fields) is that often, proposed explanations of phenomena fail to respect the difference between people and their brains. Moreover, as we have been at pains to describe, proponents of multiple-systems explanations often take suggestive, but rather inconclusive evidence at one level (e.g., brain structure) to ‘support’ accounts presented at

another level (e.g., functional organization). This combination of explanations gives rise to an illusory veneer of ‘convergent’ evidence.

If one steps back and considers the evidence at each level in isolation, the seductive appeal of multiple-system accounts diminishes markedly. At the brain level there is considerable debate about the reliability and strength of the conclusions that can be drawn from, for example, imaging studies about the operation of distinct systems (e.g., Coltheart, 2006; Page, 2006; Uttal, 2001). At the functional level, the dichotomisation of processes and representations into separable systems is muddled considerably by empirical demonstrations that are odds with the putative characteristics of the systems (see Sections 2 and 3).

Furthermore, as noted by Keren and Schul (2009), explaining phenomena via the invocation of multiple (most often dual) systems might ‘feel’ productive simply because it serves a basic cognitive function of wanting to structure and order our knowledge (i.e., ironically, *categorization*). But a good classification scheme combined with a ‘good story’ should not – though it often does – increase our confidence that an explanation is correct. It is this, in our opinion misplaced, sense of confidence that leads Poldrack and Foerde (2008) to offer the following (startling) evaluation:

“...in the end the single- versus multiple-systems approaches must be evaluated on the basis of both their explanatory power and their productivity, in terms of generating new and interesting results. We would argue that the MMS [multiple-memory-systems] view has continued to provoke new and interesting results across human and animal research, whereas the single-system view has largely focused on protecting its central assumptions and attacking results from the MMS approach rather than inspiring novel findings. It is this kind of

productivity that we believe argues strongly in favor of the MMS approach in comparison to the single-system approach.” (p.202)

However, good explanations must not only be productive or fruitful in the sense of “inspiring novel findings” (as emphasised by Poldrack and Foerde) but must also satisfy criteria like coherence, empirical accuracy, scope, internal consistency, simplicity, precision of predictions and formalism (Brewer, Chinn, & Samarapungavan, 1998; Keren & Schul, 2009). As the preceding sections highlight, many multiple-system theories fall short when considering such criteria.

### **6.3 Final thoughts**

Our goal in the present article has not been to protect assumptions, attack results and to be generally unproductive. Rather, it has been to highlight fundamental problems we perceive in currently popular multiple-systems accounts of category learning. In so doing, we have aimed to elucidate underlying assumptions, present a balanced interpretation of results, and to suggest an alternative productive program based on the methodological and conceptual analyses we have offered. Because psychological theory is a very difficult project, we need to be as certain as possible about how we interpret evidence and the manner in which we construct our explanatory accounts. For the reasons that we have outlined, we believe that our theoretical understanding of category learning is far from settled. The challenge for researchers in categorization (and many other areas) is to escape the shackles of multiple-system ‘explanations’ and to develop models and measurement procedures that will facilitate a systematic exploration of the body of data before us.

**Acknowledgements**

The support of the Australian Research Council (Grant DP: 0877510 awarded to the three authors) is gratefully acknowledged.

## References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409–429.
- Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M.A. & Poldrack, R. A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology*, 2, 1144-1152. doi: 10.1152/jn.01209.2003
- Ashby, F. G., Paul E.J., & Maddox, W.T. (in press). COVIS. In E.M. Pothos & A.J. Wills (Eds.). *Formal approaches in categorization*. New York: Cambridge University Press.
- 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, 442-481. doi: 10.1037/0033-295X.105.3.442
- Ashby, F. G. & Ell, S. W. (2002). Single versus multiple systems of learning and memory. In H. Pashler & J. Wixted (Eds.) *Stevens' handbook of experimental psychology (3<sup>rd</sup> Ed.)*, Vol 4: *Methodology in experimental psychology* (pp.655-691). Hoboken, NJ, US: John Wiley & Sons Inc.
- Ashby, F. G., Ell, S. W. & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, 31, 1114-1125.
- Ashby, F. H. & Ennis, J. M. (2006). The role of the basal ganglia in category learning. In B. H. Ross (Ed.) *The psychology of learning and motivation: Advances in research and theory (vol 46)* (pp. 1-36). San Diego: Elsevier Academic Press.
- Ashby, F. G. & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149-178. doi: 10.1146/annurev.psych.56.091103.070217

- Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M. & Ell, S. W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology*, *17*, 115-124. doi: 10.1037/0894-4105.17.1.115
- Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137-181.
- Bennett, M.R. & Hacker, P.M.S. (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell Publishing.
- Brewer, W.F. Chinn, C.A., & Samarapungavan, A. (1998). Explanation in scientists and children. *Minds and Machines*, *8*, 119–136.
- Bussemeyer, J. R., & Jones, L. E. (1983). The analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*, 549-562.
- Channon, S., Jones, M. & Stephenson, S. (1993). Cognitive strategies and hypothesis testing during discrimination learning in Parkinson's disease. *Neuropsychologia*, *31*, 175-82. doi: 10.1016/0028-3932(93)90082-B
- Coltheart, M. (1985). Cognitive neuropsychology. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and Performance* (Vol. 11, pp. 3-37). Hillsdale, NJ: Erlbaum.
- Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323-331.
- Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91-101.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgement, and social cognition. *Annual Review of Psychology, 59*, 255-278. doi: 10.1146/annurev.psych.59.103006.093629
- Filoteo, J. V. & Maddox, W. T. (2007). Category learning in Parkinson's Disease. In *Research Progress in Alzheimer's Disease and Dementia* (Ed. M.K. Sun). Vol 3. (pp 2-26). Nova Science Publishers, Inc.
- Filoteo, J. V., Maddox, W. T. & Davis, J. D. (2001). A possible role of the striatum in linear and nonlinear categorization rule learning: Evidence from patients with Huntington's disease. *Behavioral Neuroscience, 115*, 786-798.
- Filoteo, J. V., Maddox, W. T., Ing, A. D., Zizak, V. & Song, D. D. (2005). The impact of irrelevant dimensional variation on rule based category learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society, 11*, 503-513. doi: 10.1017/S1355617705050617
- 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. doi: 10.1037/0894-4105.19.2.212
- Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 103*, 11778-11783. doi: 10.1073/pnas.0602659103
- Foerde, K., Poldrack, R. A. & Knowlton, B. J. (2007). Secondary-task effects on classification learning. *Memory & Cognition, 35*, 864-874.
- Gigerenzer, G. & Reiger, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin, 11*, 23-26. doi: 10.1037/0033-2909.119.1.23

- Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- 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. doi: 10.1101/lm.45202
- Heathcote, A., Brown, S.D., & Prince, M. (in press). The design and analysis of state-trace experiments. *Psychological Methods*.
- Heffernan, M. (2009). An examination of the processes underlying probabilistic category learning. *Unpublished PhD thesis*. School of Psychology, University of New South Wales.
- Jahanshahi, M., Wilkinson, L., Gahir, H., Dharminda, A. & Lagnado, D. A. (2010). Medication impairs probabilistic classification learning in Parkinson’s disease. *Neuropsychologia*, 48, 1096-1103. doi: 10.1016/j.neuropsychologia.2009.12.010
- Keren, G. & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533-550. doi: 10.1111/j.1745-6924.2009.01164.x
- Kinder, A. & Shanks, D. R. (2003). Neuropsychological dissociations between priming and recognition: A single-system connectionist account. *Psychological Review*, 110, 728-744. doi: 10.1037/0033-295X.110.4.728
- Knowlton, B. J. (1999). What can neuropsychology tell us about category learning? *Trends in Cognitive Sciences*, 3, 123-124. doi: 10.1016/S1364-6613(99)01292-9

- Knowlton, B. J., Mangels, J. A. & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*, 1399-1402. doi: 10.1126/science.273.5280.1399
- 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*, 22-44.
- Kruschke, J. K. (2005). Category Learning. In: K. Lamberts and R. L. Goldstone (Eds.), *The Handbook of Cognition*, pp. 183-201. London: Sage.
- Kruschke, J. K. (2008). Models of categorization. In: R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology*, pp. 267-301. New York: Cambridge University Press.
- Lackner, J. R. (2009). Hans-Lukas Teuber: A remembrance. *Neuropsychology Review*, *19*, 4-7.
- Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*, 162-183. doi: 10.1037/0096-3445.135.2.162
- Loftus, G. R, Dillon, A. M, & Oberg, M. A. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*, 835-865.
- Lehrer, R. (2009). *The Decisive Moment: How the Brain Makes up its Mind*. Text Publishing: Melbourne, Australia.
- Lewandowsky, S., & Farrell, S. (in press). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage.
- Love, B. C., Medin, D. L. & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-322.

- Lovibond, P. F. & Shanks, D. R. (2002). The role of awareness in Pavlovian condition: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, 28, 3-26. doi: 10.1037/0097-7403.28.1.3
- Maddox, W. T. & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes*, 66, 309-332. doi: 10.1016/j.beproc.2004.03.011
- Maddox, W. T., Ashby, F. G. & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 650-662. doi: 10.1037/0278-7393.29.4.650
- Maddox, W. T., Ashby, F. G., Ing, A. D. & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32, 582-591.
- 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, 100-107. doi: 10.1037/0278-7393.31.1.100
- Maddox, W. T., Love, B. C., Glass, B. D. & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, 108, 578-589. doi: 10.1016/j.cognition.2008.03.010
- Marr, D. (1982). *Vision*. H. Freeman and Co.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.

- Minda, J. P. & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Ed.) *The Psychology of learning and motivation: Advances in research and theory (Vol 52)* (pp. 117-162). San Diego: Academic Press.
- Mitchell, C.J., De Houwer, J., & Lovibond, P.F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32, 183-198.  
doi:10.1017/S0140525X09000855
- Newell, B. R. & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, 12, 285-290. doi:  
10.1016/j.tics.2008.04.009
- Newell, B. R., Dunn, J. C. & Kalish M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, 38, 563-581. Doi: 10.3758/MC.
- Newell, B. R., Lagnado, D. A. & Shanks, D. R. (2007). Challenging the role of implicit processes in probabilistic category learning. *Psychonomic Bulletin & Review*, 14, 505-511.
- Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., Mesulman, M.-M. & Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, 17, 37-43. doi:  
10.1093/cercor/bhj122
- Nomura, E. M. & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience and Biobehavioral Reviews*, 32, 279-291. doi: 10.1016/j.neubiorev.2007.07.006

- Nosofsky, R. A. & Kruschke, J. K. (2002). Single system models and interference in category learning: Commentary on Waldron & Ashby (2001). *Psychonomic Bulletin & Review*, 9, 169-174.
- Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53–79.
- Nosofsky, R. A., Stanton, R. D. & Zaki, S. R. (2005). Procedural interference in perceptual classification: Implicit learning or cognitive complexity? *Memory & Cognition*, 33, 1256-1271.
- Nosofsky, R. M. & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, 9, 247-255. doi: 10.1111/1467-9280.00051
- Page, M. P. A. (2006). What can't functional neuroimaging tell the cognitive psychologist? *Cortex*, 42, 428-443. doi: 10.1016/S0010-9452(08)70375-7
- Palmeri, T. J. & Flanery, M. A. (2002). Memory systems and perceptual categorization. In B. H. Ross (Ed.). *The psychology of learning and motivation: Advances in research theory*, (vol 41) (pp.141-189). San Diego: Academic Press.
- Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C. & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, 414, 546-550. doi: 10.1038/35107080
- Poldrack, R. A. & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, 32, 197-205. doi: 10.1016/j.neurobiorev.2007.07.007

- Poldrack, R. A. & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, *41*, 245-251. doi: 10.1016/S0028-3932(02)00157-4
- Price, A. L. (2005). Cortico-striatal contributions to category learning: Dissociating the verbal and implicit systems. *Behavioral Neuroscience*, *119*, 1438-1447. doi: 10.1037/0735-7044.119.6.1438
- Price, A. L. (2009). Distinguishing the contributions of implicit and explicit processes to performance of the weather prediction task. *Memory & Cognition*, *37*, 210-222. doi: 10.3758/MC.37.2.210
- Price, A., Filoteo, J. V. & Maddox, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*, 1213-1226. doi: 10.1016/j.neuropsychologia.2009.01.031
- Ransmayr, G., Bitschnau, W., Schmidhubereiler, B., Berger, W., Karamat, E., Poewe, W., et al. (1990). Slowing of high-speed memory scanning in Parkinsons-disease is related to the severity of Parkinsonian motor symptoms. *Journal of Neural Transmission-Parkinsons Disease and Dementia Section*, *2*(4), 265-275.
- Reber, P. J., Knowlton, B. J. & Squire, L. R. (1996). Dissociable properties of memory systems: differences in the flexibility of declarative and nondeclarative knowledge. *Behavioral Neuroscience*, *110*, 861-871. doi: 10.1037/0735-7044.110.5.861
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382-407.
- Schmitt-Eliassen, J., Ferstl, R., Wiesner, C., Deuschl, G. & Witt, K. (2007). Feedback-based versus observational classification learning in healthy aging and Parkinson's disease. *Brain Research*, *1142*, 178-188. doi: 10.1016/j.brainres.2007.01.042

- Shanks, D. R. & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367-447.
- Sherry, D. F. & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*, 439-454. doi: 10.1037/0033-295X.94.4.439
- Shohamy D., Myers, C. E., Grossman, S., Sage, J. & Gluck, M. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain: A Journal of Neurology*, *127*, 851-859. doi: 10.1093/brain/awh100
- Speekenbrink, M., Channon, S. & Shanks, D. R. (2008). Learning strategies in amnesia. *Neuroscience and Biobehavioral Reviews*, *32*, 292-310. doi: 10.1016/j.neubiorev.2007.07.005
- Speekenbrink, M., Lagnado, D. A., Wilkinson, L., Jahanshahi, M. & Shanks, D. R. (2010). Models of probabilistic category learning in Parkinson's disease: Strategy use and the effects of L-dopa. *Journal of Mathematical Psychology*, *54*, 123-136. doi: 10.1016/j.jmp.2009.07.004
- Stanton, R. D. & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: Evidence against the multiple-learning systems hypothesis. *Memory & Cognition*, *35*, 1747-1758.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science* *103*, 677-680. doi:10.1126/science.103.2684.677

- Swainson, R., SenGupta, D., Shetty, T., Watkins, L. H. A., Summers, B. A., Sahakian, B. J., Polkey, C. E., Barker, R. A. & Robbins, T. W. (2006). Impaired dimensional selection but intact use of reward feedback during visual discrimination learning in Parkinson's disease. *Neuropsychologia*, *44*, 1290-1304.  
doi:10.1016/j.neuropsychologia.2006.01.028
- Tenenbaum, J. B. & Griffiths, T. L. (2001) Generalization, similarity and Bayesian inference. *Behavioral & Brain Sciences*, *24*, 629-640.
- Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology*, *6*, 267-296.
- Tharp, I. J. & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): further data demonstrate complexities in the assessment of information-integration category learning. *Cognition*, *111*, 410-414. doi: 10.1016/j.cognition.2008.10.003
- Trigg, J. & Kalish, M. (2009). Explaining how the mind works: on the relation between cognitive science and philosophy. *Manuscript submitted for publication*.
- Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458. doi: 10.1126/science.7455683
- Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732-749.
- Waldron, E. M. & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category systems. *Psychonomic bulletin & Review*, *8*, 168-176.

Wilkinson, L., Lagnado, D. A., Quallo, M. & Jahanshahi, M. (2008). The effect of feedback on non-motor probabilistic classification learning in Parkinson's disease.

*Neuropsychologia*, 46, 2683-2695. doi: 10.1016/j.neuropsychologia.2008.05.008

Zeithamova, D. & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, 34, 387-398.

## Footnotes

1. PD patients also appear to be less impaired on PCL from the outset if they are not on medication at the time of testing – a situation which is atypical of the majority of studies (Jahanshahi, Wilkinson, Gahir, Dharminda, & Lagnado, 2010).
2. The dual-tasks used by Newell et al. (2007) and Foerde et al. (2007) differed considerably – the former used a ‘numerical Stroop task’ whereas the latter used a tone-counting task. Further research is required to establish why these two tasks might have different effects on PCL (see Heffernan, 2009).
3. Note that the weather prediction task, described in Section 2 is technically a probabilistic information integration task because the optimal strategy involves integration of information from multiple cues.
4. Although the necessary contrast can be couched in terms of one or two “systems”, we do not commit ourselves to that interpretation and argue later for an alternative conceptualization.
5. These data were derived from Table 2 of the published article. This table presents the means and standard errors for each condition which included three levels of time interval (2.5s, 5s, and 10s). The data in Figure 6 have been averaged over this factor.
6. These data were obtained from Figures 4 and 7 of Filoteo et al (2001) and Figure 4 of Filoteo et al (2005).
7. We illustrate these problems in relation to state-trace analysis. Exactly the same issues apply to the interpretation of dissociations.
8. As well as to the incorrect inference that the two tasks could be dissociated.

9. The critical nature of these predictions is that what others (e.g., Medin & Schaffer, 1978) have identified as the role of dimensional attention in categorization are actually the result of a dimension-dependent rule-based system competing with an exemplar system. The identification of dimensional attention with explicit, verbalizable rules would suggest that no non-verbal creature could show attention learning, which is obviously false.

10. We will discuss in Section 6 whether this notion of an algorithm is unproblematic. If it is meant to imply that these computations happen 'in the mind' then it is; if it is meant to imply that these computations describe the functional organization of the capacity, then it is not.

## Figure Captions

*Figure 1.* The “Weather Prediction Task” – the most commonly used variant of a Probabilistic Category Learning (PCL) Task. The task requires participants to predict the weather on the basis of different cue configurations. Each card has a different predictive validity. Two cards are “strong” predictors of one or other outcome (e.g., the triangle card predicts ‘rain’ with 0.8 probability, and the square card predicts ‘sun’ with 0.8 – note in the diagram only validities with respect to rain are indicated). The other two cards are ‘weak’ predictors, with either 0.4 or 0.6 predictive validity. In the standard ‘feedback’ version of the task participants make predictions and are given trial-by-trial corrective feedback. In the alternative ‘observation’ version cues and the outcome appear together on each trial and participants are asked to learn (memorize) the cue-outcome pairings.

*Figure 2.* Top panel: Mean importance ratings averaged across the two strong and the two weak cues in a feedback version of the weather prediction task. Note that values on the y-axis represent how much participants said they relied on each card, with 4 = “Greatly”, 3 = “Moderately”, 2 = “Slightly”, 1 = “Not at all”. Bottom panel: Mean implicit regression weights (absolute values) for strong and weak cards derived from conducting rolling regressions across a moving window of trials. Note in both figures blocks are 10 trials. (Adapted from Figure 11 of Lagnado, Newell, Kahan, & Shanks, 2006).

*Figure 3.* (A): Distribution of stimuli in a rule-based (RB) category structure. (B): Distribution of stimuli in an information integration (II) category structure. (C) An example of a Gabor patch stimulus. In a typical experiment a single stimulus is shown on the screen and participants have to categorize it as an “A” or “B” and then receive corrective feedback.

*Figure 4.* Schematic diagrams of the effect of number of learning trials and delay between response and feedback on performance on rule-based (RB) and information integration (II) tasks. (a) According to COVIS. (b) According to a single-system model.

*Figure 5.* Example state-trace plots. (a) State-trace plot consistent with COVIS and other multiple-system models. (b) State-trace plot consistent with the single-system model shown in Figure 4b. The numbers on each axis refer to the percentage of correct responses on the information integration (II) and rule-based (RB) tasks with each data point corresponding to performance averaged across one of four blocks of learning trials.

*Figure 6.* Observed state-trace plot averaged over delay interval (from Maddox, Ashby & Bohil, 2003). The numbers on each axis refer to the percentage of correct responses on the information integration (II) and rule based (RB) tasks with each data point corresponding to performance averaged across an 80 trial learning block. Error bars correspond to standard errors of the mean.

*Figure 7.* Combined state-trace plot from Filoteo, Maddox & Davis (2001) and Filoteo, Maddox, Salmon & Song (2005). HD refers to Huntington's disease patients and NC (HD) refers to the corresponding group of normal controls (data from Figures 4 and 7 of Filoteo et al, 2001). PD refers to Parkinson's disease patients and NC (PD) refers to the corresponding group of normal controls (data from Figure 4 of Filoteo et al, 2005).

Figure 1

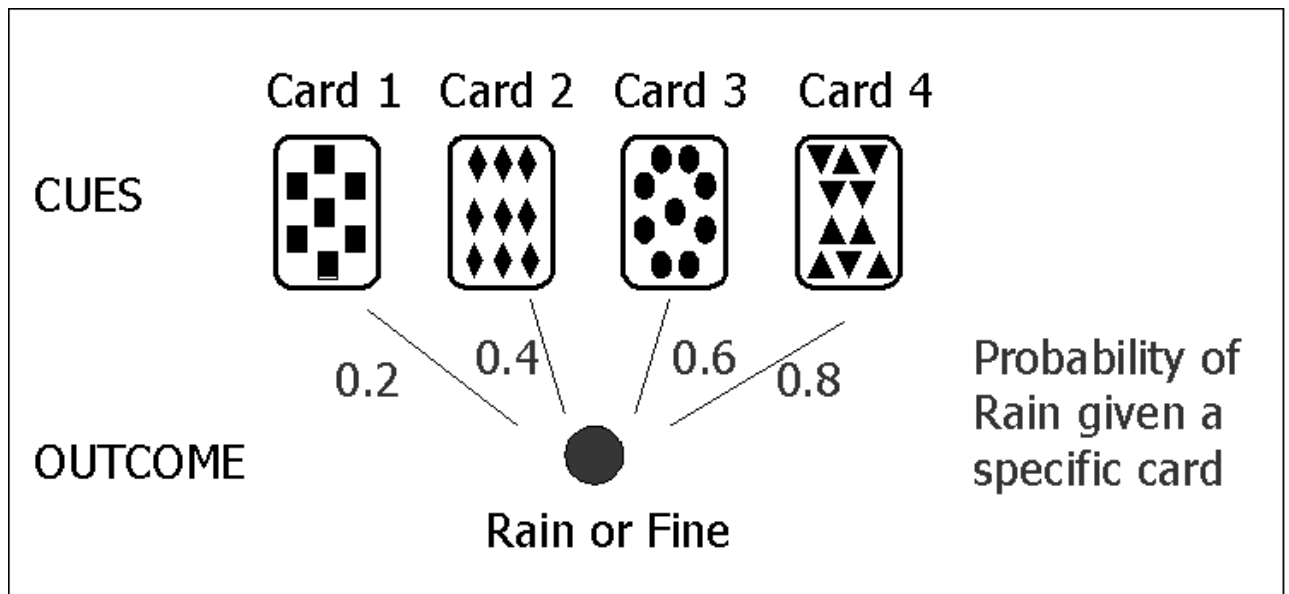


Figure 2

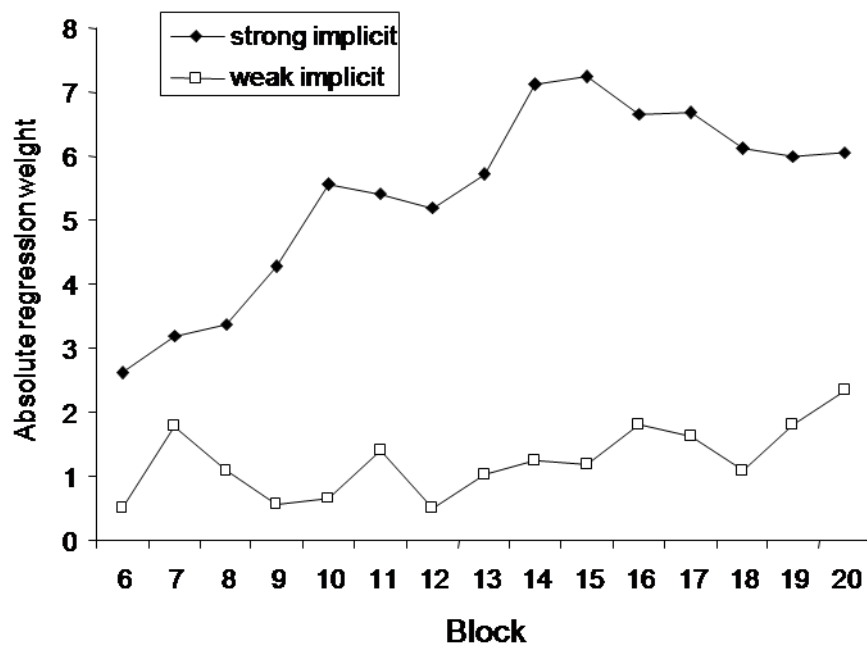
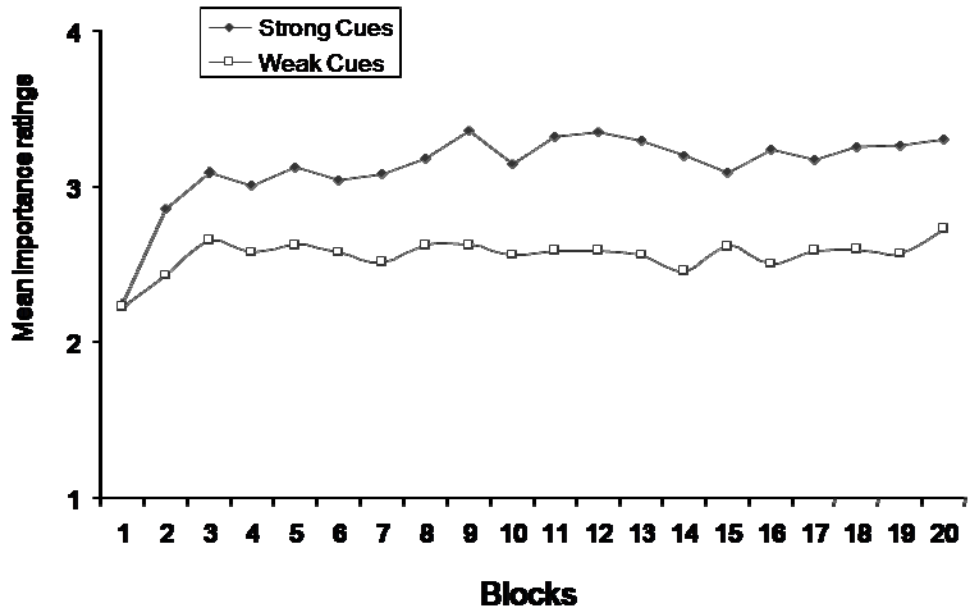


Figure 3

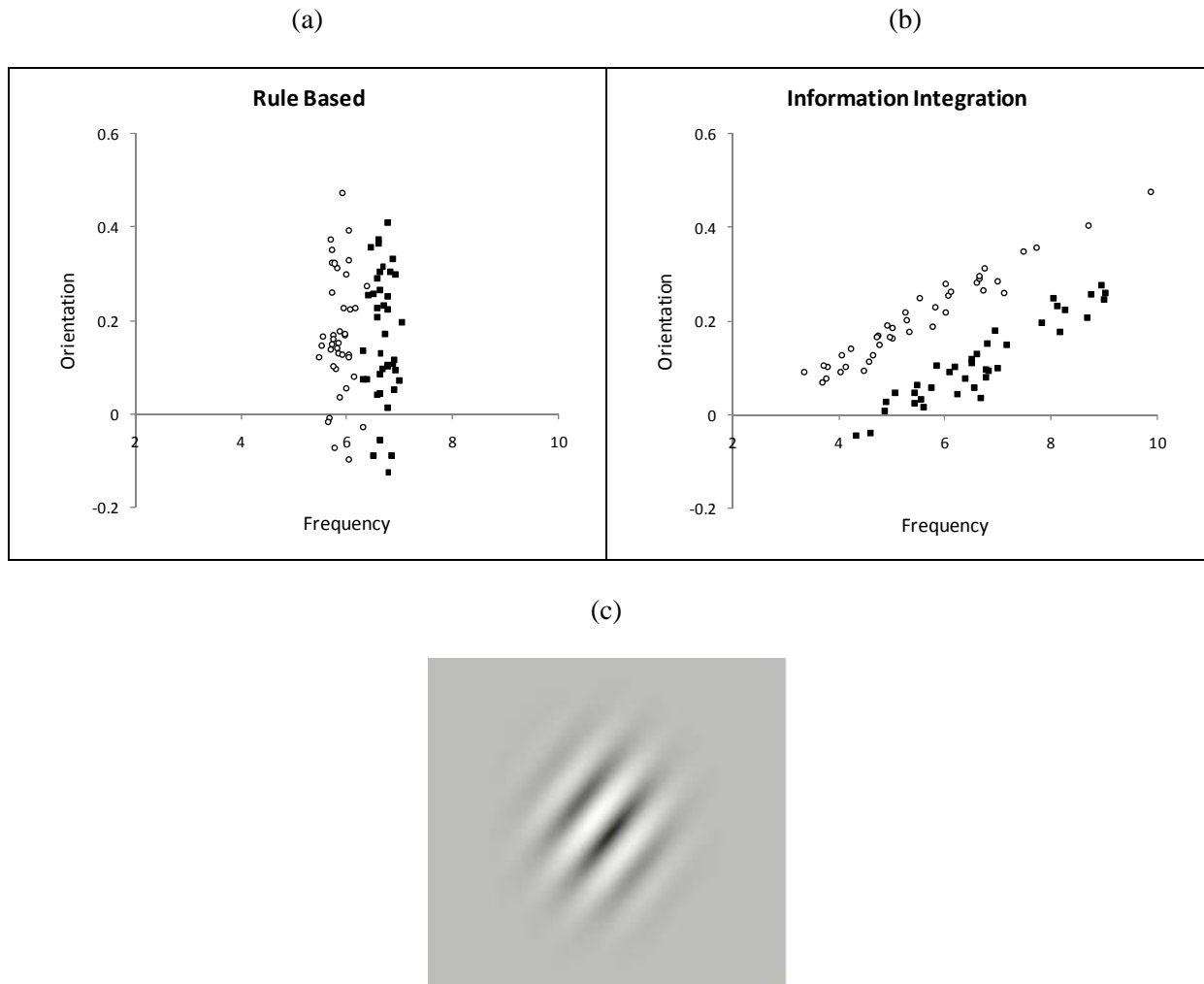
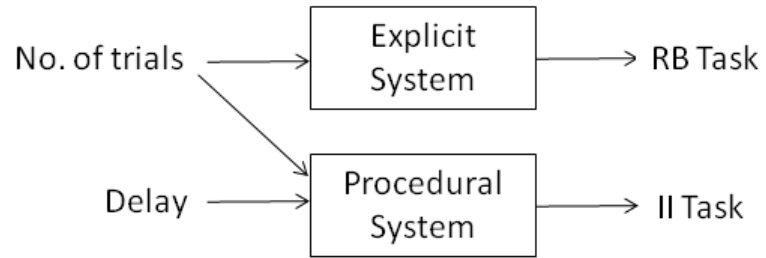


Figure 4

(a)



(b)

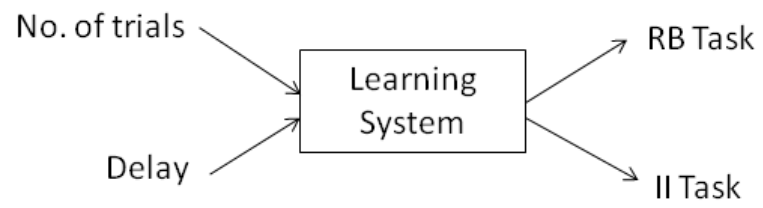


Figure 5

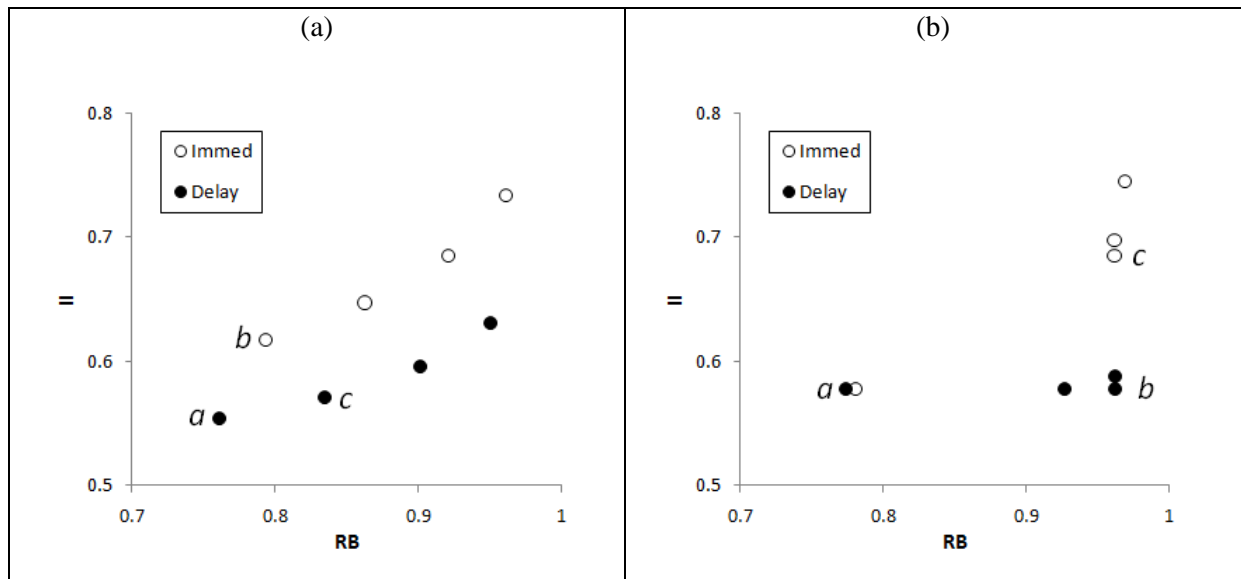


Figure 6

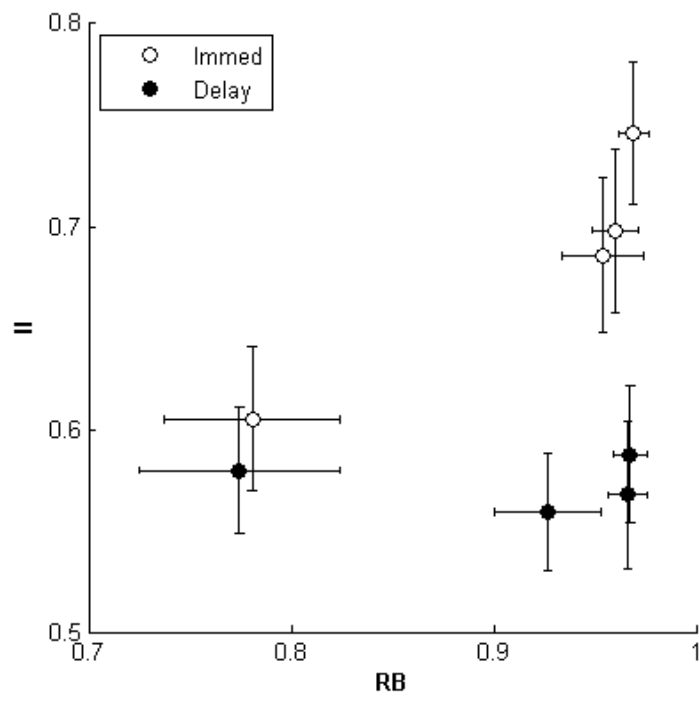


Figure 7

