2017

# Critique of a dual-system model of category learning

Edmunds, C E R

http://hdl.handle.net/10026.1/8817

http://dx.doi.org/10.24382/648
University of Plymouth

# CRITIQUE OF A DUAL-PROCESS MODEL OF CATEGORY LEARNING

by

**CHARLOTTE E R EDMUNDS**

A thesis submitted to Plymouth University in partial fulfilment for the degree of

**DOCTOR OF PHILOSOPHY**

School of Psychology

**September 2016**

# Acknowledgements

First and foremost, I would like to thank Andy Wills for his superb supervisor skills. This project would have been half as good and even less fun without him. Thank you for working with me and long may it continue.

Special thanks also go to Fraser Milton for his support, attention to detail and sharing of resources. A large proportion of the experiments presented here would not exist without his insight, logical rigour or participants points. Thank you.

I would like to thank the attendees of Learning Lunch for their cheerful skepticism as well as their terrifying (but thought-provoking) questions.

I would like to thank Angus Inkster and Gemma Williams for their help in coding the verbal reports given by participants. I would also like to thank Gemma for her assistance in running Experiment 10, as well as her endless patience with my organisational skills.

I would like to thank my friends Nick Lange, Marie-Claire Reville, Martina Hlavata, Yong Chen and all those at EMG Symphony Orchestra for their support (most often displayed as a never-ending supply of coffee). I would also specifically like to thank Marion Wood for saying, in an early discussion about this work where I was complaining that I didn't know what the participants were doing, "Well, why don't you ask them?" For this single comment (and for everything else) thank you.

Finally, I would like to thank my parents, Tom and Sue Edmunds. My words cannot convey how grateful I am for your support and love. So instead, as promised, this work is dedicated . . .

*To my parents and my father's theory of dinosaurs*

# Authors declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. Work submitted for this research degree at Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

Relevant scientific conferences were regularly attended at which work was presented. Three papers have been accepted for publication in refereed journals or conference proceedings.

Word count for the main body of this thesis: **49,296**

**Signed:** _____

**Date:** _____

**Publications:**

**Edmunds, C. E. R. & Wills, A. J. (2016)** Modelling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

**Edmunds, C. E. R., Wills, A. J. & Milton, F. (2016)** Memory for exemplars in category learning. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

**Edmunds, C. E. R., Milton, F. & Wills, A. J. (2015)** Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology, 68(2),*1203–1222.

**Posters and conference presentations:**

**38th Annual Conference of the Cognitive Science Society** Philadelphia, USA, 2016. Conference presentation: *Memory for exemplars in category learning.* Edmunds, C. E. R., Milton, F. & Wills, A. J.

**38th Annual Conference of the Cognitive Science Society** Philadelphia, USA, 2016. Poster presentation: *Modelling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS.* Edmunds, C. E. R., & Wills, A. J.

**Associative Learning Symposium XVIII** Gregynog, Wales, 2016. Poster presentation:

*Comparison of recognition memory for exemplars between rule-based and information-integration category structures.* Edmunds, C. E. R., Wills, A. J. & Milton, F.

**International Convention of Psychological Science (ICPS)** Amsterdam, Holland, 2015. Invited talk: "Dual-system approaches to cognition: Some thoughts and concerns." Edmunds, C. E. R. & Wills, A. J, in symposium chaired by Visser, I. & Jansen, B. titled "Criteria and validity of dual systems models in cognition."

**Associative Learning Symposium XVII** Gregynog, Wales, 2015. Conference presentation: *"Initial training with difficult items does not facilitate information-integration category learning"* Edmunds, C. E. R. & Wills, A. J.

**Experimental Psychology Society Meeting** London, England, 2015. Conference presentation: *"Feedback is superior to observational training for both rule-based and information-integration category learning"* Edmunds, C. E. R. & Wills, A. J.

**Associative Learning Symposium XVII** Gregynog, Wales, 2014. Conference presentation: *Oral presentation: "Feedback is superior to observational training for both rule-based and information-integration category learning"* Edmunds, C. E. R., Milton, F. N., & Wills, A. J.

# Abstract

Some researchers have argued that the category learning literature is conclusive: people learn to generalise from past experiences to novel ones using multiple learning systems (Ashby & Maddox, 2011). To the extent that this claim is true, it is due in no small part to work investigating the predictions of the dual-system model COVIS (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998). The work presented here investigates the evidence for this model.

In Chapter 1, I describe the main features of the COVIS model and briefly review some of the evidence that supports this model. This section highlights two main problems with this literature. First, that many of the studies that have purported to support a dual-systems description of category learning have been found to be flawed when re-examined by independent researchers. This observation is explored more deeply in the next two chapters, where I attempted to reproduce the work of two studies argued to support the COVIS literature. In Chapter 2, I re-examined the work reported by Ashby, Maddox, and Bohil (2002) that looked at the effect of training type on category learning. In Chapter 3, I re-examined work reported by Spiering and Ashby (2008) that looked at the effect of training order on category learning. Both these chapters failed to find evidence for two systems of category learning.

The second issue raised in Chapter 1 is that none of the studies cited in support of COVIS critically examined the fundamental assumptions of the model. More specifically, Chapters 4 and 5 looked at how participants complete the categorisation tasks used in this literature. In Chapter 4, I conducted experimental work to determine whether some category structures are learned implicitly, as argued throughout the COVIS literature. Then, in Chapter 5 I conducted several simulations to investigate an analysis used ubiquitously in the COVIS literature to determine the strategies participants use to complete categorisation tasks. This analysis is a critical manipulation check for all these experiments. However, I found evidence that the analysis systematically over-estimates the evidence for dual-systems. Furthermore, both these chapters found evidence to suggest that the evidence for the COVIS model can be explained without needing to assume that participants can learn implicitly.

Finally, in Chapter 6 I bring these threads together to discuss the implications for the COVIS model.

# Contents

# List of Figures

# Chapter 1

# Introduction

Some researchers have argued that the category learning literature is conclusive: people learn to generalise from past experiences to novel ones using multiple learning systems (Ashby & Maddox, 2011; Ashby & Valentin, 2016). To the extent that this claim is true, it is due in no small part to work investigating the predictions of the dual-system model COVIS (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Paul, & Maddox, 2011). The current work focuses on investigating the experimental evidence supporting this model in two ways: first, by re-examining several key COVIS-supporting studies and second, by investigating the validity of two critical experimental assumptions within this literature.

## 1.1 The COVIS model

The COVIS model proposes that category learning is mediated by two functionally and biologically independent systems that compete to control responding (Ashby et al., 1998, 2011). It assumes that people begin learning how exemplars correspond to category labels using a process of hypothesis testing. This process is mediated by the Verbal System. This system implements explicit, verbalisable rule-based strategies using declarative memory and selective attention.* The Verbal System implements rules as a discriminant function. First, one or more stimulus dimensions are selected to use to categorise the stimulus (for example, the size of the stimulus). Then, the stimulus is compared to a criterion on those dimensions to determine whether it is low or high on that dimension (for the rule based on size, then this defines “small” and “big”). Then, each of these attributes is associated with a category label, resulting in a rule that, for the size dimension, might be described verbally as “small stimuli in Category A and large stimuli in Category B.” If this rule does not result in high accuracy, the participant might switch rules to see whether

*Note that throughout the current work I discriminate between “strategies” and “rules.” Here, strategies are any type of systematic approach to sorting stimuli into categories. Rules are a special kind of strategy that divide the stimulus space using boundaries that are parallel and/or perpendicular to the stimulus dimensions.

other stimulus dimensions are better at predicting category membership. Biologically, this system is hypothesised to be mediated by the frontal and medial temporal lobes (Nomura et al., 2007; Ashby & Valentin, 2016).

If the category structure is difficult to verbalise, COVIS assumes the Verbal System cannot learn it adequately and so the Procedural System[†] will take control of responding (Ashby et al., 1998, 2011). This system can learn a greater variety of category structures than the Verbal System by using a procedural learning mechanism driven by perceptual similarity. A stimulus is hypothesised to activate a unit corresponding to that stimulus, as well as units of stimuli that appear similar to it. The stimulus will be predicted to be in Category A if it, and the stimuli perceptually similar to it, have previously been associated with Category A. If this results in an incorrect response, the association of the stimulus with that category label is adjusted using reward prediction error. The Procedural System is hypothesised to result in "category knowledge that is opaque to declarative consciousness" (Smith et al., 2015, p. 2476). Biologically, it is hypothesised to be mediated by neural connections between the perceptual and motor systems situated at the striatum (Nomura et al., 2007; Ashby & Valentin, 2016).

COVIS assumes that switching between Verbal and Procedural Systems is controlled by a Competition mechanism. To decide which system will guide responding, this mechanism uses a combination of the "confidence" each system has in its response, and the level of "trust" the competition system has in each system. Confidence is conceptualised as how certain each system is in its predictions for the category membership for the stimulus on that trial. For example, the Verbal System would be very confident in its categorisation decision if the stimulus was very far from the criterion value (for example tiny or huge, if the rule was based on size). On the other hand, the Procedural System would be very confident if the stimulus was very similar to stimuli of one category and very dissimilar to stimuli of the other category. The level of trust for each system is conceptualised as how often that system has been right in the past. COVIS assumes that initially the level of trust in the Verbal System is very high. Then, trust is adjusted downwards every time the participant makes a mistake. The trust in the Procedural System is determined from the trust in the Verbal System: as trust in the Verbal System goes down trust in the Procedural System goes up. Recent work has proposed there is a hard switch from

[†]Note that I refer to this system as the Procedural System, rather than the "Implicit System" referred to in the acronym, to avoid later confusion when discussing whether or not this system is implicit or explicit.

the Verbal System to the Procedural System; once a participant has switched to the Procedural System, they do not go back to using the Verbal System (Paul & Ashby, 2013).

## 1.2 Details of a representative COVIS experiment

One of the strengths of COVIS (compared to other models in the category learning literature) is that it describes the processes of category learning in three ways (Lewandowsky, Palmeri, & Waldmann, 2012): formally (in terms of mathematics), psychologically (in terms of behaviour) and biologically (in terms of brain areas or properties; Ashby et al., 1998, 2011). This facilitates research as it allows the predictions from one description of the model to constrain predictions of the others (Wills & Pothos, 2012). That being said, the core of evidence for COVIS comes from functional dissociations found in behavioural experiments (for reviews see Ashby & Maddox, 2005, 2011). These experiments look at how an experimental manipulation affects the learning of two kinds of category structure: rule-based and information-integration (Crossley, Paul, Roeder, & Ashby, 2015; Ashby & Valentin, 2016). These two category structures are hypothesised to be learned by the Verbal and Procedural Systems respectively (Smith et al., 2015).

Rule-based category structures can be described using Boolean logic (Feldman, 2003), which corresponds to decision boundaries that are parallel or perpendicular to the stimulus dimensions (Ashby & Valentin, 2016). The rule-based category structure most often used in the COVIS literature is a unidimensional one (Figure 1.1a). This structure sorts stimuli based on a single stimulus dimension, such as “small stimuli are in Category A and large stimuli are in category B.” In stimulus space, this corresponds to a category boundary perpendicular to the stimulus dimension of interest (in this example, stimulus size). Another type of rule-based category structure occasionally used in this literature is a conjunction (Figure 1.1b). This structure sorts stimuli using a logical AND rule, such as “if stimuli are small and dark they are in Category A, else they are in Category B.” In stimulus space, this corresponds to partitioning a quarter of the space using one line parallel to a particular dimension and another perpendicular to that dimension. COVIS assumes that rule-based category structures, like the unidimensional and conjunction structures, are optimally learned by the Verbal System. This is because the category structures are easy to describe verbally and so rule-based strategies implemented by the Verbal System can score highly.

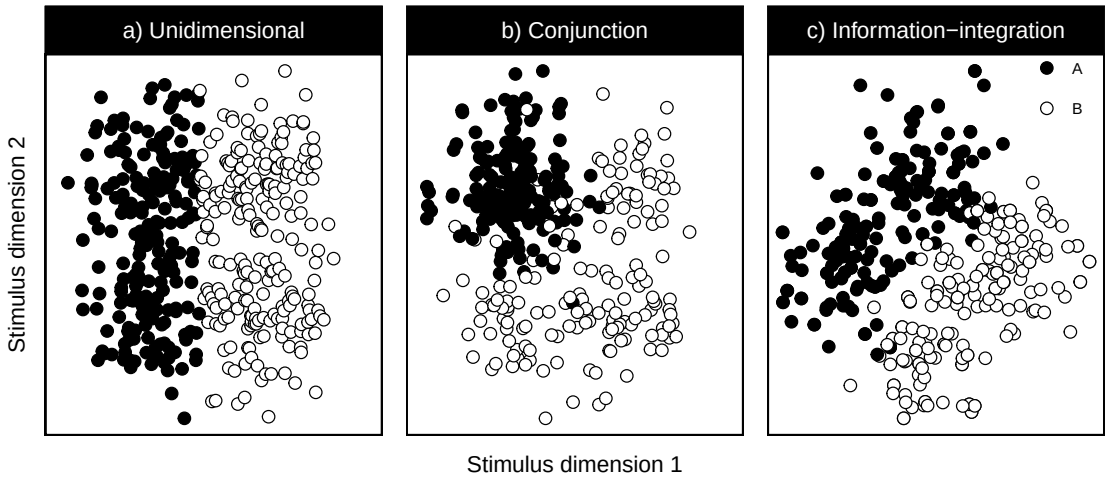


*Figure 1.1:* Examples of the most common category structures used in the COVIS literature represented in stimulus space. Each point represents a single stimulus, black points would be in one category and white points would be in the other.

The other kind of category structure used in the COVIS literature is the information-integration category structure (Figure 1.1c). In stimulus space, this structure corresponds to a diagonal boundary, i.e. a boundary not parallel to either of the stimulus dimensions. Critically, this structure is difficult to describe verbally as it requires participants to make the stimulus dimensions commensurable (Ashby & Valentin, 2005). This means COVIS predicts that this category structure is learned by the Procedural System, because verbalisable rule-based strategies implemented by the Verbal System will not be successful.

A particular experiment is argued to support the COVIS model if there is a dissociation between the effect of an experimental manipulation on rule-based and information-integration category learning. The dissociation can be of three different kinds (Ashby & Valentin, 2016). First, the manipulation can disrupt learning of the rule-based category structure but not learning of the information-integration category structure (e.g. Zeithamova & Maddox, 2006). Second, the manipulation can disrupt learning of the information-integration category structure but not the rule-based category structure (e.g. Maddox, Ashby, & Bohil, 2003). Third, the experiment might find evidence of a double dissociation: the manipulation may improve learning of one category structure but hinder learning of the other (e.g. DeCaro, Thomas, & Beilock, 2008). These category structure dissociations have been argued to support the existence of two qualitatively different systems of category learning as they imply that there are underlying mental functions required by rule-based category learning but not information-integration category learning and vice versa (Dunn & Kirsner, 2003).

## 1.3 Issues identified with the COVIS evidence

There is a large amount of experimental evidence based on this type of dissociation cited in support of the COVIS model (Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2016). However, the strength of evidence for COVIS depends not on the number of studies, but rather the number of these that are shown to be reliable and valid. Indeed, there is a growing body of work that casts doubt on the validity or interpretation of a large proportion of these experiments (for a partial review see Newell, Dunn, & Kalish, 2011). In the following, I briefly review this literature.

One issue identified within the COVIS literature is that, in some experiments, not all of the of the participants have adequately learned the category structure they were assigned. This results in those experiments comparing rule-based and information-integration category structure conditions confounded by the proportion of non-learners in each condition. Newell, Dunn, and Kalish (2010) demonstrated that this was the case in experiments conducted by Zeithamova and Maddox (2006). In these experiments, participants learned rule-based and information-integration category structures with or without a concurrent load. They found that concurrent load reduced rule-based learning but not information-integration category structure learning, thereby supporting the COVIS model. However, when Newell et al. repeated the experiment and removed the participants that failed to complete the concurrent task well and those that were guessing in the category learning tasks, the dissociation disappeared. Furthermore, they they found a one-dimensional state-trace, which is indicative of a single-system of learning (although see Ashby, 2014; Yeates, Wills, Jones, & Mclaren, 2015).

Tharp and Pickering (2009) made similar objections to two other experiments purported to support the COVIS model. Waldron and Ashby (2001) found that a concurrent load reduced learning of a rule-based category structure but had no impact on information-integration learning. Using a similar task, DeCaro et al. (2008) looked at the effect of individual differences in working memory capacity on category learning performance. They found that participants with superior working memory capacity performed better in the rule-based task, but worse in the information-integration task. However, Tharp and Pickering pointed out the learning criterion used in these studies meant that participants could reach the criterion whilst still not understanding the category structure. Indeed, when DeCaro, Carlson, Thomas, and Beilock (2009) replicated their original study using

the more stringent learning criterion proposed by Tharp and Pickering, they no longer found a double dissociation between working memory capacity and category learning task. Rather, participants with larger working memory capacities learned both rule-based and information-integration category structures better. This finding is supported by other work using structural equation modelling that demonstrated performance in both rule-based and information-integration category structures is correlated with working memory capacity (Lewandowsky, Yang, Newell, & Kalish, 2012).

Other critiques have identified confounds in some of the studies argued to support COVIS. For example, Filoteo, Lauritzen, and Maddox (2010) investigated the effect of interrupting working memory processes on category learning. On each trial, they interrupted working memory by using a secondary working memory task immediately after the participant received feedback on their categorisation response. They found that the secondary task harmed learning of a rule-based category structure and improved learning of an information-integration one. This result was argued to support the COVIS model because the rule-based category structures are learned using the Verbal System that relies on working memory processes. Therefore, interrupting working memory processing would have a large negative effect on learning mediated by the Verbal System. Additionally, it would facilitate the switch to the Procedural System thereby improving learning of the information-integration category structure. However, Newell, Moore, Wills, and Milton (2013) pointed out that the participants given the secondary task also had an extra 2500ms to explicitly consider the feedback for the category learning task. When Newell et al. manipulated the length of the inter-trial interval and the position of the working memory task, they found that participants with a smaller inter-trial interval performed worse than those with a longer inter-trial interval, whether or not the working memory task was immediately after training. This is easily explained using a single-system of category learning, and so does not provide additional evidence for the COVIS model.

The most substantial portion of the critiques of the COVIS experiments relates back to the heavy reliance of this literature on dissociation logic. This logic is inherently difficult to interpret and comes with several criteria that must be met before a dissociation can be taken as evidence for dual-systems of category learning (Dunn & Kirsner, 2003). For example, for a difference in average performance to validly support the COVIS model, it must be true that the category structures differ in no other way than in the verbalisability of the optimum boundary (Newell & Dunn, 2008; Newell et al., 2010). Otherwise, the differ-

ential effect of the manipulation on performance in the two category tasks may be due to the extraneous differences between the tasks and cannot be attributed to the underlying mechanisms. One example of this was identified by Stanton and Nosofsky (2007) who re-examined two experiments by Maddox, Ashby, Ing, and Pickering (2004). Maddox, Ashby, et al. looked at the effect of interrupting feedback processing on rule-based and information-integration category learning. They found that adding a working memory task immediately after feedback impaired rule-based but not information-integration category learning. This effect is consistent with the COVIS model as it assumes that the Verbal System relies on working memory processes, whereas the Procedural System does not. Therefore, inserting a working memory task immediately after feedback should impair rule-based but not information-integration category learning. However, Stanton and Nosofsky demonstrated that this effect was actually due to reduced perceptual discriminability between the categories in the rule-based category structure compared to the information-integration category structure. They found that a rule-based category structure with increased discriminability was not negatively impacted by interrupting working memory, and that an information-integration category structure with decreased discriminability was. Therefore, the original effect can be attributed to differences in discriminability between the category structures, not participants using two distinct systems of category learning.

Another critique demonstrated that the sensitivity of the information-integration task to additional perceptual noise was responsible for another dissociation. Maddox et al. (2003) examined the effect of feedback delay on unidimensional rule-based and information-integration category learning. They compared a condition where participants were told whether or not they were correct immediately after responding with conditions where the feedback was displayed 2.5, 5 or 10 seconds after responding. They found that varying feedback delay did not affect rule-based category learning, but that performance was poorer in the delayed conditions than the immediate condition for participants learning the information-integration category structure. This is consistent with the COVIS model as it assumes that changing feedback properties, such as feedback timing, should have very little impact on Verbal System learning as it is hypothesised to rely on working memory. This means the Verbal System can maintain a representation of the stimulus until the feedback arrives. Therefore, participants should be able to learn rule-based category structures under a variety of circumstances, even if the feedback occurs after a long

delay or sometimes even without feedback at all (Ashby, Queller, & Berretty, 1999; Ell, Ashby, & Hutchinson, 2012). In contrast, proponents of COVIS assume that changing feedback properties should have a large impact on Procedural System learning. This is because the Procedural System relies on a chemically mediated reward signal to learn. In other words, the Procedural System relies on local learning mechanisms based in the striatum. This means that learning in the Procedural System requires a close temporal relationship between the stimulus, response and feedback. Any delay in feedback should therefore disrupt learning.

However, as noted by Maddox and Ing (2005), there were several problems with the design of Maddox et al. (2003) because the category structures were not matched for the number of relevant dimensions for optimal performance. In the unidimensional rule-based category structure, the participants only need to pay attention to one of the stimulus dimensions, whereas for the information-integration category structure optimal responding involves attending to both stimulus dimensions. Therefore, it may be that feedback delay matters for learning both category structures, but, because the unidimensional category structure is easier, the effect is not visible. To overcome this objection Maddox and Ing (2005) replicated the experiment using different category structures. These category structures contain four categories in a cross configuration with the category boundaries either parallel to the stimulus dimensions (rule-based condition) or diagonally offset from the stimulus dimensions (information-integration condition). Here again, they found that varying feedback delay made no difference to learning of the rule-based category structure but performance was poorer with delayed feedback than with immediate feedback for information-integration learning.

In contrast, Dunn, Newell, and Kalish (2012) argued that the differential effect of feedback delay on learning these category structures was due to the mask chosen by Maddox and colleagues (2003, 2005). In these experiments, after the Gabor patch stimulus was presented another Gabor patch was displayed as a mask. When Dunn et al. replaced this with a pattern mask, the effect disappeared. Dunn et al. postulated that this was due to the fact that adding the Gabor patch mask after viewing a Gabor patch stimulus interrupts processing (Magnussen, 2000), thereby increasing perceptual noise. COVIS fails to explain this finding as it postulates that any delay should have an effect on information-integration category learning, not just delay after a Gabor patch mask.

Dunn et al. (2012) make a similar argument to explain the findings by Maddox, Love, Glass, and Filoteo (2008). This study looked at the effect of feedback type, either minimal or full. With minimal feedback, participants are only told whether they are correct or incorrect. With full feedback, participants are told whether or not they are correct as well as the correct category label. Maddox et al. found that, compared to minimal feedback, full feedback improves rule-based category learning but harms information-integration category learning. They hypothesised that this was because full feedback leads to more rapid rule learning. This would help learning of the rule-based structure by the Verbal System. On the other hand, it would also lead to inappropriate sustained reliance on the Verbal System when learning the information-integration category structure, thus harming performance. However, Dunn et al. again found that their ability to replicate the pattern of results reported by Maddox et al. (2008) was dependent on the type of mask used.

In addition to structural differences between the category structures, for a dissociation to validly support COVIS the rule-based and information-integration category structure tasks must be of similar difficulty. These arguments have been made several times. For example, Ashby, Ell, and Waldron (2003) trained participants on either a rule-based or information-integration category task. Then, in a transfer phase, some participants were asked to switch the buttons they used to indicate their categorisation response. Switching buttons had a detrimental effect on information-integration category learning, but not rule-based category learning. Ashby, Ell, and Waldron argued that this was consistent with COVIS, as the information-integration category structure is learned by the Procedural System, which maps stimuli to response locations. However, Nosofsky, Stanton, and Zaki (2005) argued that this was because the rule-based task was not sensitive enough to demonstrate an interruption to the procedural learning component present in both rule-based and information-integration category structure learning. When they reduced the amount of time participants had to respond, thus making the rule-based test more sensitive, rule-based category learning was also harmed by a response switch (although see Maddox, Lauritzen, & Ing, 2007, for a response).

Similarly, Zaki and Kleinschmidt (2014) demonstrated that the differential effect of response mapping on category learning found in several studies (Maddox, Bohil, & Ing, 2004; Hélie & Ashby, 2012) was also due to differences in cognitive complexity between categorisation tasks. For example, Maddox, Bohil, and Ing (2004) looked at rule-based and information-integration category learning with either a consistent or inconsistent re-

sponse mapping. In the consistent response mapping condition, participants pressed a button corresponding to the category they thought the stimulus belonged to ('A' or 'B'). In the inconsistent response condition, participants pressed a button corresponding to the answer to one of two questions ('Is this an A?' or 'Is this a B?') resulting in an inconsistent response mapping for a particular stimulus. They found that an inconsistent mapping had a detrimental effect on information-integration category learning but not on rule-based category learning. They argued this was consistent with the COVIS model, as information-integration category structures are optimally learned by the Procedural System that maps perceptions to response locations. However, Zaki and Kleinschmidt demonstrated that this effect was better explained by differences between the number of relevant stimulus dimensions between categorisation tasks. Maddox, Bohil, and Ing used a unidimensional rule-based category structure, which reduces the cognitive load of the task. When Zaki and Kleinschmidt used a bi-conditional rule-based task, which is also predicted to be learned by the Verbal System, the participants also demonstrated superior performance in the consistent response mapping condition compared to the inconsistent response mapping condition.

Another requirement for making valid inferences from a dissociation is that all the participants in each category structure condition are of a single type (Dunn & Kirsner, 2003). For example, this assumption may be violated if there are different proportions of non-learners in each cell of the experiment (as discussed above; Newell et al., 2010). Another possibility that would result in this type of difference is that the number of participants using the optimum strategy for each category structure may differ (this is discussed in much greater detail in Chapter 5). To avoid this objection, proponents of COVIS typically use a model-based analysis informed by General Recognition Theory (GRT; Ashby & Townsend, 1986; Ashby & Soto, 2015, 2016) as a manipulation check. GRT is a multidimensional version of signal detection theory (Macmillan & Creelman, 2005). Briefly, this approach assumes that participants' strategies can be modelled as (usually linear) decision boundaries that pass through stimulus space. If the decision boundary matches the category boundary in the category structure, then the participant is assumed to be using the optimum strategy for the category structure. If the majority of participants in each condition is found to be using the optimal strategy, then any functional dissociations found are argued by the proponents of COVIS to demonstrate the existence of two systems of category learning. However, Donkin et al. (2015) found that the proportion

of participants found to be using the optimal strategy critically depends on the category structure and other features of the strategy analysis. This indicates that less faith should be taken in the conclusions of the COVIS literature.

The failure to control for extraneous differences between the category structure conditions is not just limited to the behavioural realm. Nomura et al. (2007) used functional magnetic resonance imaging (fMRI) to see which brain areas were related to rule-based and which to information-integration category learning. Their results were consistent with COVIS. However, Carpenter, Wills, Benattayallah, and Milton (2016) repeated this experiment with better matched category structures and found the opposite. Contrary to the predictions of COVIS, they found that information-integration category learning was associated with larger medial temporal lobe activation than rule-based category learning.

This brief review of the COVIS literature highlights two problems. First, the evidence for the COVIS model is not as conclusive as portrayed by its supporters. Several reviews of the COVIS literature from these authors have failed to cite any substantial critiques of the model at all (Ashby & Maddox, 2011; Ashby & Valentin, 2016). A different picture emerges from a more neutral review of the literature. On closer examination, a substantial amount of COVIS-supporting studies have been found to be invalid, confounded or more easily described by single-system approaches to category learning. This large number of critiques means that it is important to re-examine the findings of those studies not yet examined (e.g. Ashby et al., 2002; Casale, Roeder, & Ashby, 2012; Smith et al., 2014, 2015; Spiering & Ashby, 2008).

This is especially important given that very few studies in psychology are ever replicated, the gold standard for scientific evidence (Ledgerwood, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). For instance, the Open Science Collaboration (2015) found that only 36 out of 97 studies replicated still demonstrated significant results. Furthermore, Makel, Plucker, and Hegarty (2012) estimated that only 1.07% of psychology studies published since 1900 were replications of previous work. Therefore, in the first half of this thesis I will be re-examining two results cited in support of the COVIS model: in Chapter 2 I re-examine the differential effect of training type on category learning found by Ashby et al. (2002) and in Chapter 3 I re-examine the differential effect of training order found by Spiering and Ashby (2008).

Second, it becomes apparent that studies within the COVIS literature tend not to eval-

uate the fundamental assumptions of the model or experimental design. The experiments within this literature assume that the COVIS model is true and that rule-based and information-integration category structures map onto the two systems of COVIS (a particularly clear example of this is Smith et al., 2015). Indeed, to my knowledge, no study within the COVIS canon has been explicitly designed to discriminate between COVIS and an alternative model of category learning. This bias towards confirmation makes it difficult to evaluate whether the COVIS model is a good description of category learning compared to other models (Nickerson, 1998; Wills & Pothos, 2012). Therefore, in the second half of the current work I examine two of these assumptions. In Chapter 4, I examine the assumption that information-integration category structures are learned implicitly. Then, in Chapter 5 I examine in more detail the model-based strategy analysis typically used in the COVIS literature as a manipulation check to determine whether participants are using the optimal strategy for the category structure condition they were assigned.

## 1.4 The structure of the thesis

This thesis has six chapters. Chapters 2 and 3 report work that re-examines two key studies argued to support the COVIS model. Chapter 2 examines the effect of observational training, compared to the more standard feedback training, on rule-based and information-integration category learning. Previous evidence within the COVIS literature has argued that observational training harms learning, but only for category structures learned by the Procedural System (Ashby et al., 2002). However, these experiments include several other differences between the category structure conditions, aside from the critical difference of verbalisability. The experiments reported in Chapter 2 show that once these extraneous factors are removed, the dissociation disappears.

Chapter 3 examines the effect of training order on learning information-integration category structures. Counterintuitively, Spiering and Ashby (2008) found that initial training on hard examples improved information-integration learning compared to initially training on easy examples. In four experiments, I demonstrate that this finding is likely a false-positive.

In Chapters 4 and 5, I report work that examines two fundamental assumptions of the experimental COVIS literature. In Chapter 4, I directly examine the assumption that information-integration category structures are learned implicitly. Four experiments examine recognition memory performance following a category training phase of either a

rule-based or information-integration category structure. Contrary to the predictions of the COVIS model, these experiments provide evidence that both types of category structure are learned explicitly using verbalisable rules.

In Chapter 5, I examine the evidence that participants learn information-integration implicitly, an assumption ubiquitously used in the COVIS literature. These studies use a model-based strategy analysis to determine whether participants were using the optimal strategy for the category structure they learned. If this is the case for participants in the information-integration conditions, they are assumed to be learning implicitly using the Procedural System. In this chapter, I argue that this role of the model-based strategy analysis is both crucial for the logic of their experiments and deeply flawed. To do this, I report the results of several model-recovery simulations and demonstrate that an experiment argued to support the COVIS model (Smith et al., 2015) is in fact explainable using a single rule-based system of learning.

In Chapter 6, I draw together the key results reported in the previous five chapters and consider the implications for the COVIS model, as well as directions that future research on these issues could take.

## 1.5 Miscellaneous notes on the text

### 1.5.1 Data archiving

The data and stimuli as well as experimental and analysis scripts will be available on publication of the relevant studies. However, these archives are available on request in the interim.

### 1.5.2 Analyses

All analyses presented here were conducted using R (R Core Team, 2015).

# Chapter 2

# Feedback type

## Notes

Note that Experiments 1 and 2 were published in Edmunds, Milton, and Wills (2015). Additionally, Experiment 1 was originally conducted as part of a Masters degree at the University of Exeter. That being said, all the analyses (aside from the null-hypothesis significance testing) reported for this experiment are novel and conducted as part of my PhD.

## 2.1 Introduction

In this chapter, I report two experiments that investigated the effect of training type on category learning. These experiments were based on work by Ashby, Maddox, and Bohil (2002) in which they examined the effects of observational and feedback training. With observational training, participants were first shown the category label, then shown the stimulus and asked to give a confirmatory response. With feedback training, participants were first shown the stimulus, then gave a classification response followed by the correct category label. The effect of training type was examined on two category structures: one unidimensional and one information-integration. Ashby et al. found that the type of training made no difference for unidimensional category learning. However, when learning the information-integration category structure performance was better with feedback training compared to observational training.

Ashby et al. (2002) argue that these results support the COVIS model because they hypothesise that rule-based and information-integration category structures are most effectively learned by dissociable neural systems that utilise feedback differently (Ashby et al., 1998, 2011). As discussed in Chapter 1, rule-based category structures, such as the unidimensional category structures used in these experiments, are hypothesised to be learned by the explicit, Verbal System. This system is hypothesised to be mediated by

working memory and so Ashby et al. hypothesised that participants using this system can maintain the representation of the category label across the whole trial. Therefore, they would not predict any differences in learning between the training type conditions when learning a rule-based category structure.

In contrast, the information-integration category structure is difficult to describe verbally and so COVIS predicts that it will be learned by the implicit, Procedural System (Ashby et al., 1998, 2011). The Procedural System relies on reward prediction error to learn. This is absent in the observational condition as participants are never uncertain about which category a particular stimulus is in. Therefore, as was found in Ashby et al. (2002), feedback training should result in superior performance to observational training.

Critically, this explanation depends on the unidimensional, rule-based category structure being easy to verbalise and the information-integration category structure being hard to verbalise. This is because it is the verbalisability of the category structure that determines which learning system should control responding. However, Ashby et al. (2002) contained three superfluous factors (aside from verbalisability) that also varied between the unidimensional rule-based and information-integration category structures. First, the number of stimulus dimensions required to optimally learn the category structures differed: the information-integration category structure required attention to both stimulus dimensions, whereas the unidimensional structure only required attention to one. Classification based on a single stimulus dimension has been shown to sometimes require fewer cognitive resources (as indexed by the effects of concurrent load and time pressure) than multidimensional classification (Milton, Longmore, & Wills, 2008; Wills, Milton, Longmore, Hester, & Robinson, 2013; Wills, Inkster, & Milton, 2015). Therefore, training type may be less critical in the rule-based conditions compared to the information-integration condition simply because the rule-based structure is less demanding to learn.

Second, participants in Ashby et al.'s (2002) first experiment made fewer errors learning the rule-based structure, compared to learning the information-integration category structure. This raises the possibility that the observed dissociation was due to a ceiling effect. Ashby et al. partially addressed this possibility in a second experiment.* There, they repeated the experiment solely on the unidimensional category structure but increased the difficulty of learning this structure by reducing the between-category difference. Although

*To avoid confusion, experiments that have been reported in the literature will always be associated with a citation, whereas those I conducted that are reported here will not.

the average accuracy did decrease, there was still no statistically significant difference in performance between training type conditions. This is consistent with Ashby et al.'s conclusions. That being said, by the end of training, feedback training did appear to be slightly superior to observational training. In addition, difficulty was only increased for one of the two combined counterbalance conditions, and there were only five participants per condition. Thus, the lack of a significant difference in Ashby et al.'s Experiment 2 might be attributable to a lack of statistical power.

Third, for both of Ashby et al.'s (2002) experiments the between-category difference is lower in the unidimensional, rule-based category structure than in the information-integration category structure. Category separation can be defined as the mean difference between-category items as plotted in stimulus space, divided by the within-category variance along the direction of the comparison. Given that differences in category separation were shown by Stanton and Nosofsky (2007) to be responsible for another purported dissociation within the COVIS literature (Maddox, Ashby, et al., 2004), it seems important to control for this in future studies of the COVIS model.

Although controlling for these factors (number of dimensions, errors and between-category separation) simultaneously whilst maintaining the key difference in verbalisability can be difficult, other category structures within the COVIS literature have achieved this goal. Specifically, Filoteo et al. (2010) in their study on the effects of interrupting working memory processing on rule-based and information-integration category learning, employed the category structures shown in Figure 2.1. Filoteo et al.'s rule-based structure is a conjunction rule that requires participants to attend to both stimulus dimensions. Furthermore, Filoteo et al.'s paper establishes empirically that these structures are well matched on average error rates. Also, the between-category distance is approximately well matched.

## 2.2 Experiment 1

### 2.2.1 Introduction

This experiment reexamined the effect of observational and feedback training on the learning of the structures used in Filoteo et al. (2010). For this experiment, COVIS predicts that feedback training should be superior to observational training, but only for participants learning the information-integration category structures. However, Ashby et al. (2002) is also consistent with the prediction that feedback training will be superior to ob-

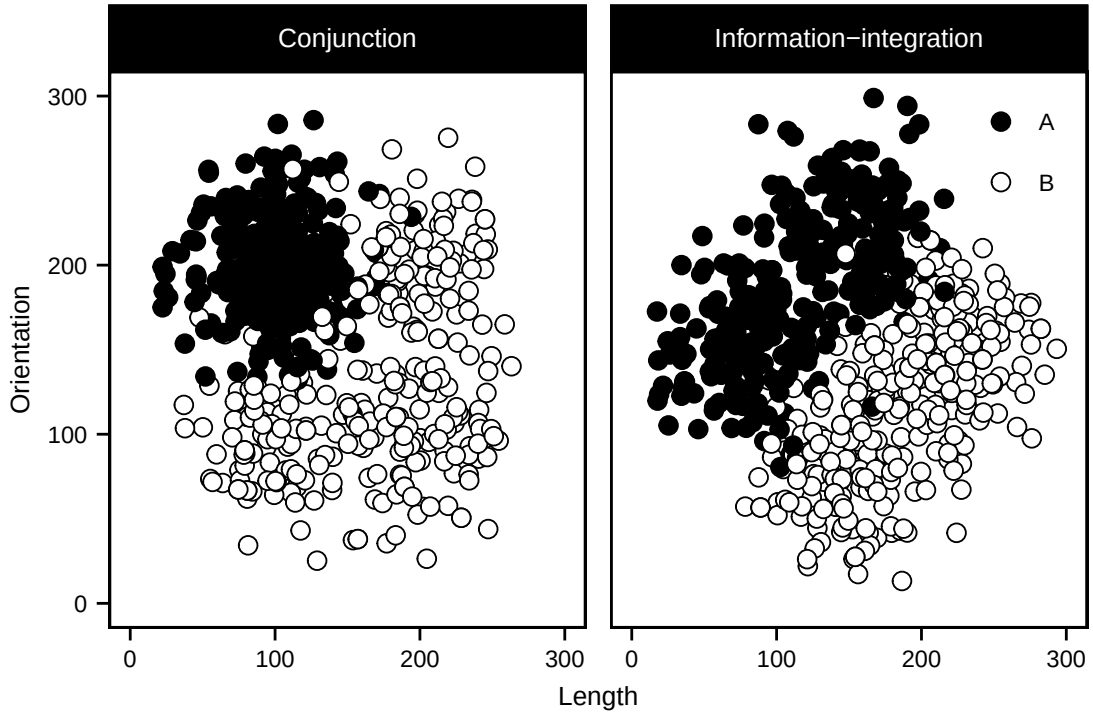


*Figure 2.1:* The category structures used in Filoteo et al. (2010) and Experiment 1 in abstract stimulus space. Each point represents a particular stimulus. Filled circles represent Category A and unfilled circles represent Category B.

servational training for both rule-based and information-integration category structures. This is because the dissociation observed by Ashby et al. might be due to one of the superfluous factors identified above, rather than the existence of two systems of category learning. If this is the case, controlling for these factors by using alternative category structures should remove the dissociation between category structures.

In addition to examining response accuracies, I also asked participants to report at the end of the experiment the strategy they used during the experiment. Not only does previous evidence indicate that reported strategy use can be informative when comparing feedback and observational training on a probabilistic category learning task (Newell, Lagnado, & Shanks, 2007), but it can also directly assess whether participants can verbalise the category structure. If the conjunction category structure is indeed easier to verbalise, then participants should be better able to describe the underlying structure in the conjunction condition than the information-integration category structure.

Also, I will perform an analysis commonly used in the COVIS literature to check that the category structure manipulation has resulted in the appropriate switch between learning systems. This analysis is based on General Recognition Theory (GRT; Ashby & Townsend, 1986; Ashby & Gott, 1988) a multi-dimensional version of signal detection

theory, and is ubiquitously used in the COVIS literature. I have some reservations about this procedure, which are detailed more fully in Chapter 5. However, for comparison with other research the analysis is included here.

### 2.2.2 Method

#### Participants

80 participants (47 female) were recruited from the University of Exeter community and were not rewarded for their participation.

#### Design

The experiment had a 2 (category structure: rule-based, information-integration) x 2 (training type: feedback, observational) between-subjects, factorial design. 20 participants were randomly assigned to each condition. Participants in the feedback condition were first shown the stimulus, then were given an opportunity to respond, followed finally by the correct category label. Participants in the observational condition were first shown the correct category label, then the stimulus, followed finally by the opportunity to respond. Category learning was measured by the proportion of correct responses in each block.

#### Stimuli

The stimuli used were those used in the two-dimensional conjunction and information-integration conditions of Filoteo et al. (2010), whose representations in abstract stimulus spaces are shown in Figure 2.1. Each stimulus was a single black line on a white background that varied on two dimensions: line length and orientation. In both conditions, maximum accuracy was 95% as 5% of the stimuli overlapped the optimal category boundary.

#### Materials

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a MacBookPro with a 15-inch screen.

#### Procedure

Participants in all conditions were informed that they would be shown a series of lines that varied in length and orientation, that their task was to assign the lines to either Category A or Category B and that approximately half the lines were in each category. They were

also told that at the beginning of the experiment they may have to guess but by the end they should be able to reach high levels of accuracy. They were further informed of the structure of the experiment, the format of the trials, the position of feedback within the trial (which varied between conditions) and the response keys.

The experiment consisted of 10 blocks of 60 trials, with 600 trials in total, separated by self-paced breaks. Participants assigned stimuli to either Category A (by pressing the 'Z' key) or Category B (by pressing the '/' key). Starting with a training block, the blocks alternated between training and test. This was to provide a measure of performance during learning for both observational and feedback conditions as well as to facilitate comparison with Ashby et al. (2002). The training trials of the feedback learning conditions consisted of displaying the stimulus for 500ms, followed by a blank screen for 500ms, followed by a self-paced classification response. Finally the correct category label was displayed for 500ms. In the observational learning condition training trials consisted of first displaying the correct category label for 500ms, followed by a blank screen for 500ms, followed by the stimulus for 500ms to which the participant made a self-paced response. The test trials in both feedback and observational training conditions included no information about the correct category assignment and consisted of a stimulus displayed for 500ms followed by a self-paced response. The intertrial interval in all conditions was 500ms.

At the end of the experiment, participants were presented with a questionnaire that asked them to describe whether they had a specific strategy when classifying the items and, if so, to describe it, using either words or pictures. They filled in their answers in a blank space.

### 2.2.3 Results

Following Ashby et al. (2002), analyses were conducted on the final test block of the data from all participants. Conducting the analyses across all test blocks lead to the same conclusions, as did excluding participants failing to reach 50% on the final block (the analysis method and exclusion criterion applied by Filoteo et al., 2010). Figure 2.2 shows the mean proportion accuracy for each condition across all test blocks and just in the last test block. The trial-level raw data for Experiment 1 are archived at `www.willslab.co.uk/exe201201`.

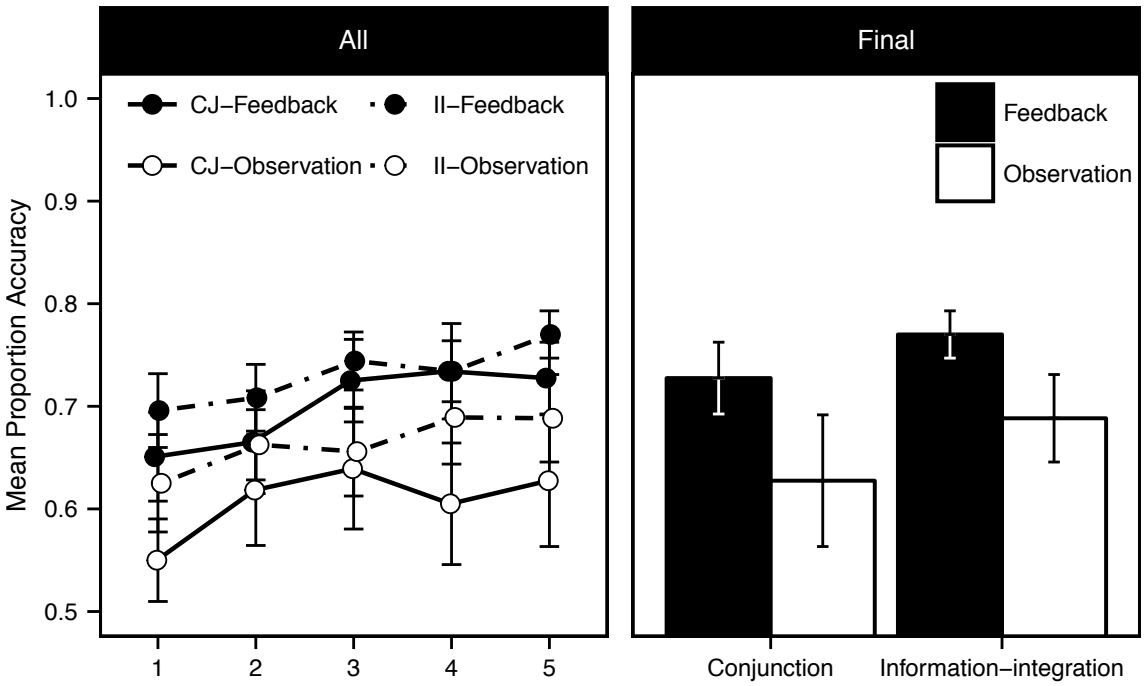


*Figure 2.2:* The results of Experiment 1. Error bars are 95% difference-adjusted confidence intervals (Baguley, 2012).

### Null-hypothesis significance testing

There was a statistically significant main effect of training type $F(1,76) = 9.4$, $\eta^2 = 0.11$, $p = 0.003$. Participants trained using feedback had higher performance, $M = 0.75$, $SD = 0.09$, than those given observational training $M = 0.66$, $SD = 0.17$. However, the effect of category structure did not reach significance $F(1,76) = 3.04$, $\eta^2 = 0.04$, $p = 0.085$. The key interaction predicted by COVIS also failed to reach statistical significance, $F(1,76) = 0.10$, $\eta^2 = 0.00$, $p = 0.758$.

### Bayesian analysis

The standard statistical analyses above indicate that, unlike in Ashby et al. (2002), there appears to be no difference between rule-based and information-integration categories in the effect of training type on learning. In other words, feedback training is superior to observational training no matter the type of category structure to be learned. However, in null-hypothesis significance testing, non-significant results are ambiguous: they could either be due to insufficient statistical power or due to the null hypothesis being correct (Dienes, 2011). As the interaction between feedback type and category structure formed the basis of the conclusions drawn by Ashby et al. (2002), it is important to determine whether the reason the current study failed to find an effect was because it lacked power.

One way of determining this is the case is to calculate Bayes Factors for the relevant

comparisons (Dienes, 2011). Briefly, if the Bayes Factor is over three then the experiment has found evidence for the experimental hypothesis whereas if the Bayes Factor is less than a third, the experiment finds evidence for the null hypothesis (Jeffreys, 1961). A Bayes Factor of one indicates that the evidence is exactly neutral with respect to the experimental and null hypotheses (Dienes, 2011). Values between a third and three are typically interpreted as indicating that the experiment was not sensitive enough and no conclusions can be drawn.

Here, I calculated Bayes Factors in R (R Core Team, 2015) using the R script implemented by Baguley and Kaye (2010) according to the procedure laid out by Dienes (2011). This requires the expected average difference between the two differences to be specified (the prior). In Ashby et al. (2002), the observed mean difference of differences between the information-integration conditions in Experiment 1 and the rule-based conditions in Experiment 2 was approximately 15%. This cross-experimental difference was used because the rule-based structure in Experiment 2 was better controlled for differences in overall error rates (similar to the current experiment). Following the recommendations of Dienes (2011), I assumed a two tailed normal distribution around this value with standard deviation of half the mean (i.e., 7.5, representing the experimental hypothesis that differences as small as zero are unlikely). These calculations result in a Bayes factor of $4.65 \times 10^{-4}$. As the Bayes Factor is less than a third, it indicates that the data provide support for the null hypothesis–that is, the effect of the training type manipulation does not differ between rule-based and information-integration category structures.

**State-trace analysis**

The analyses above indicate that there are no differences between the final acquisition of rule-based and information-integration categories. However, these analyses do not consider the qualitative pattern of learning throughout the experiment. As the key conceptual claim of COVIS is that there are two mechanisms of learning, it could be argued that these analyses have failed to identify multiple systems only because the difference in learning between training types just happened to be the same for rule-based and information-integration learning by the end of training. To examine the validity of this claim, I used state-trace analysis (Bamber, 1979; Loftus, Oberg, & Dillon, 2004), which has previously been used with great success on this type category learning data (Newell et al., 2010; Dunn et al., 2012).

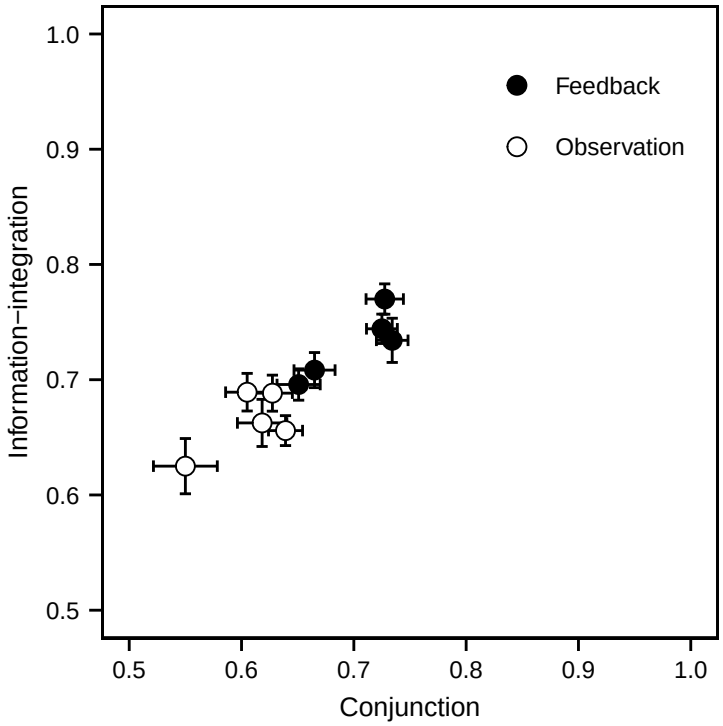


*Figure 2.3:* State-trace plot for Experiment 1 with rule-based and information-integration performance on each block on the axes. Error bars are one standard error.

State-trace analysis is an alternative to dissociation logic that allows experimenters to determine whether multiple systems are required to explain an experimental result. This is accomplished by drawing a state-trace plot. To do this, two dependent variables, such as performance on the rule-based or information-integration category structure in this case, are plotted on the $x$ and $y$ axes. Then, a trace is plotted for each training type condition, with each point being the accuracy from each test block. The state-trace plot is then inspected to determine whether the traces are consistent with a single- or multiple-system account. If the two traces overlap to form a single monotonic function, then there is an absence of evidence that a multiple-process account is required to explain the observations. If the traces form two monotonic functions, then this is often interpreted as being more supportive of a multi-process account, although the question of what the term "multi-process" means in the context of state-trace analysis has been the topic of recent debate (Dunn, Kalish, & Newell, 2014). In brief, both Yeates and colleagues (Yeates, Wills, Jones, & McLaren, 2012; Yeates et al., 2015) and Ashby (2014) have identified situations where models typically considered to be single-system accounts can produce two functions on a state-trace plot through variation in a single parameter (specifically, learning rate in the Simple Recurrent Network, Elman, 1990, and attention weight in the Generalized Context Model, Nosofsky, 1986).

From visually inspecting Figure 2.3, the data from the current experiment forms a single monotonic curve. This once again suggests an absence of evidence that a multi-process account such as COVIS is required to account for the current results.

**Model-based strategy analysis**

The predictions of COVIS for this experiment are contingent on the assumption that the category structure manipulation corresponds to a change in the strategies that participants use, and thus, the learning system that controls responding (Ashby & Maddox, 2005, 2011). Practically, this assumption is widely applied in the COVIS literature as meaning that there should be more people using rule-based strategies in the rule-based category conditions than in the information-integration conditions. Similarly, there should be more participants using a diagonal strategy in the information-integration category structure conditions than in the rule-based conditions.

Experimental studies within the COVIS framework utilise model-based analysis constructed from GRT (General Recognition Theory; Ashby & Townsend, 1986; Ashby & Gott, 1988) to examine this assumption. For each participant, this analysis determines the optimum decision boundary in stimulus space that separates the stimuli judged by each participant to be in Category A from those in Category B. Each participant is then assigned a strategy type on the basis of characteristics of their optimum boundary. The assumption that the category type manipulation has resulted in a change of category learning system is argued to be valid if more participants are using the optimum decision bound for the category structure they have been assigned to, such as a diagonal decision boundary in the information-integration conditions, than are using that strategy in the inappropriate category structure, such as a diagonal decision boundary in the rule-based conditions.

The model-based strategy analysis determines which of a pre-defined set of decision-boundary models best describes the classification responses each participant produced. The set of models considered were those used in Ashby et al. (2002) and were as follows:

The *unidimensional* models assume that the participant determines a criterion along one of the stimulus dimensions, either orientation or length. They then make a decision about the category membership of each stimulus by comparing the appropriate stimulus attribute with the criterion value. As an example, for length, this corresponds to a rule of the type: ‘Assign to Category A if the stimulus is long, or Category B if short’. The unidimensional models have two parameters: the value of the criterion and the variance of

internal (criterial and perceptual) noise.

The *conjunction* model assumes that the participants make two judgements, one for each stimulus dimension, and then combine these to make a judgement about category membership. The conjunction rule in the current analysis was of the type: 'Assign to Category A if the stimulus is short and upright, otherwise assign to Category B'. The conjunction model had three parameters: the two criterion values and internal noise.

The *general linear classifier* (GLC) model assumes that the decision boundary between the categories can be described by a straight line that can vary in gradient and intercept. The unidimensional models are therefore special cases of the GLC model. The GLC model has three parameters: the intercept and slope of the decision bound, plus noise. Unlike Ashby et al. (2002), an idealised diagonal model with a single parameter was not used in the information-integration category structure conditions. This was to prevent biasing the model-based analysis towards finding diagonal strategies in those conditions.

The *random* model assumes that participants are responding randomly; it has no parameters.

For each participant, the best fit of each of these models was calculated, and the best-fitting model selected using Akaike's Information Criterion (AIC; Akaike, 1974), which is defined as

$$AIC = -2\log L + 2p \tag{2.1}$$

where $L$ is the maximum likelihood for a particular model and $p$ is the number of parameters in the model.

The results from this analysis, which was performed using the `grt` package in the R environment (Matsuki, 2014), are reported in Table 2.1. Within the COVIS framework, the unidimensional and conjunction models are considered to represent explicit, rule-based strategies, while the diagonal (GLC) strategy is considered to represent an implicit, information-integration strategy.

In ordinal terms, the results of this analysis are consistent with the intended effects of the experimental manipulation, as seen through the lens of the COVIS model and GRT-based model analysis. Specifically, the proportion of participants best fit by a conjunction model is higher in the rule-based condition than the information-integration condition,

*Table 2.1:* The proportion of participants in each condition fit by each GRT model.

| Condition | Strategies | | | | |
|---|---|---|---|---|---|
| | CJ | UDO | UDL | GLC | RND |
| Overall | | | | | |
| Rule-based | 0.375 | 0.25 | 0.125 | 0.175 | 0.075 |
| Information-integration | 0.225 | 0.375 | 0.075 | 0.325 | 0 |
| By condition | | | | | |
| Rule-based | | | | | |
| Feedback | 0.35 | 0.3 | 0.15 | 0.15 | 0.05 |
| Observation | 0.4 | 0.2 | 0.1 | 0.2 | 0.1 |
| Information-integration | | | | | |
| Feedback | 0.15 | 0.45 | 0.1 | 0.3 | 0 |
| Observation | 0.3 | 0.3 | 0.05 | 0.35 | 0 |

Strategies: CJ=Conjunction, UDO=Unidimensional strategy based on stimulus orientation, UDL=Unidimensional strategy based on stimulus length, GLC=General linear classifier, RND=Random.

and the proportion of participants best fit by the diagonal (GLC) model is higher in the information-integration condition than in the rule-based condition.

It is perhaps not particularly surprising that some participants are best fit by a unidimensional model, as a single-dimension strategy can optimally achieve approximately 75% accuracy in both the rule-based and the information-integration conditions. From a COVIS perspective, it is not particularly problematic if some participants in the rule-based condition are in fact employing a unidimensional rule, as this is still a rule-based strategy and readily verbalizable. It is potentially more problematic from a COVIS perspective that there is a reasonable proportion of participants best fit by unidimensional models in the information-integration conditions, potentially implying the presence of significant rule-based responding in these conditions. That being said, a similar result was observed in Ashby et al. (2002), although the proportion is higher in the current study.

The presence of unidimensional responders in an information-integration condition is typically accommodated within COVIS by assuming that some participants have not yet transitioned from the Verbal System to the Procedural System. The lower proportion of participants best fit by unidimensional models in Ashby et al. (2002) may be due to the fact that Ashby et al., in their modelling of their information-integration condition, constrained the diagonal (GLC) model to have the gradient and intercept defined by the category structure. This constrained version of the model has just one parameter, while the uncon-

strained version I employed has three parameters. In an AIC model-selection procedure, reducing the number of free parameters of a model will, other things being equal, increase the proportion of participants best fit by that model. Somewhat surprisingly, Ashby et al. state that they employed the unconstrained version of the diagonal (GLC) in their fits of their rule-based condition. This difference in fitting procedure between experimental conditions seems odd, and may have contributed to the higher proportion of unidimensional classifiers in their rule-based conditions compared to their information-integration conditions.

In summary, the model-based procedures that are standard in this field broadly support the supposition that participants in the rule-based conditions classify the stimuli differently to participants in the information-integration conditions. The fact that a conjunction model best fits more participants in the rule-based condition than the information-integration condition, and a diagonal (GLC) model best fits more participants in the information-integration condition than the rule-based condition, is broadly consistent with the predictions of the COVIS model. Of course, what is not consistent with the COVIS model is that, despite these differences, there is no difference in the size of the feedback advantage in the rule-based and information-integration conditions.

Although seldom reported within the COVIS literature (although see Ashby & Vucovich, 2016, for a Bayesian approach), it is also informative to look at the performance of the best-fitting model relative to the performance of the competing models. If the winning model performs much better than its competitors, we can be more confident that this model provides the best description of the participant's behaviour, from among the pre-specified alternatives. On the other hand, if the competing models perform almost as well as the winning model, our confidence that the winning model provides the best description should be lower.

One principled way of evaluating the model-based analysis is by calculating the normalised probability that a conjunction model is preferred to the diagonal (GLC) model for each participant (or vice versa). This is done by calculating the Akaike weight, $w_i(AIC)$, for each model for each participant (Wagenmakers & Farrell, 2004). This is defined as the probability that model $i$ is the best, in terms of minimising the AIC, given the data and the set of competing models. From the Akaike weights, the normalized probability that

Model $i$ is to be preferred over Model $j$ is calculated using

$$\frac{w_i(AIC)}{w_i(AIC) + w_j(AIC)} \tag{2.2}$$

where $w_i(AIC)$ and $w_j(AIC)$ are the Akaike weights for models $i$ and $j$ respectively.

For the rule-based category structure conditions the probability of the 'best' model being a conjunction is 0.635 in the feedback training condition and 0.668 in the observational training condition. This provides additional support that participants are genuinely using orthogonal decision boundaries to make decisions. However, for the information-integration category structures the probability of the best model being the general linear classifier is much lower: 0.297 for the feedback training condition and 0.382 for the observational training condition. Clearly, confidence in the results of GRT-based model fitting in the information-integration conditions should be low.

**Verbal report analysis**

An alternative explanation of these findings from within the COVIS framework might be that the majority of participants in both the rule-based and information-integration category structure conditions were using the Procedural System. It is possible that by increasing the number of relevant dimensions in the rule-based structure, participants found this too difficult and so resorted to using the Procedural System. To investigate this possibility I examined the strategies reported by participants as summarised in Table 2.2.

The verbal reports were independently coded by myself and AJW. Any discrepancies that were not due to human error were easily resolved through discussion. First, each verbal report was examined to determine whether the participant had reported an explicit categorisation strategy or not. The inter-rater reliability for this was perfect, $\kappa = 1$, $p < .001$. Second, the available strategy descriptions were sorted into groups of three main kinds: unidimensional, two-dimensional and miscellaneous.

Participants were placed in the *unidimensional length* or *unidimensional orientation* groups if they described categorising stimuli based solely on line length or line orientation respectively.

Participants were placed in the *conjunction* group if they used both stimulus dimensions and described categorising stimuli using a logical conjunction rule such as 'short, upright lines were in Category A, otherwise they were in Category B.'

Participants were placed in the *information-integration* group if they described attempting to make the stimulus dimensions commensurable, such as 'Stimuli for which the line was longer than it was upright should be assigned to category A' or if they said anything that could be reasonably interpreted as a statement that they based their classification on overall similarity. Note that overall similarity descriptions are commonly found in other studies, outside of the COVIS-framework, which have elicited verbal reports (e.g., Wills et al., 2013).

Participants were placed in the *two-dimensional* group if they described using both stimulus dimensions but with descriptions that were too unclear to be assigned to more specific categories.

All remaining participants were assigned to the *other* group, which included participants whose descriptions were too vague to be assigned to another group.

Inter-rater reliability for strategy assignment was high, $\kappa = .813$, $p < .001$, with the majority of discrepancies appearing to be due to human error in applying the strategy definitions, rather than any inherent ambiguity in the definitions themselves (as all discrepancies were rapidly resolved by reference to the strategy descriptions). There were no significant differences between all conditions in the number of participants who did not report a strategy, $\chi^2(1) = 0.12$, $p = .730$. With respect to the types of strategy reported, there are very different patterns of responding between the rule-based and information-integration category structure conditions. For the rule-based conditions, although there is clearly some variability, the modal strategy correctly described the conjunction structure. In addition, none of the participants in these conditions reported using an overall similarity or information-integration strategy, and only 20.1% reported using unidimensional strategies.

In contrast, no participant in the information-integration category structure conditions reported any strategy that could be interpreted as describing the structure of the information-integration category they had been presented. In these conditions, participants were equally likely to report a unidimensional strategy as they were to report a conjunction rule, although strategies employing both dimensions were the majority indicating a sensitivity to the fact that both dimensions were relevant. This summary is supported by the fact that the number of participants in the rule-based category conditions who reported the optimal strategy for the categorisation problem they had been presented (44.8% of the

*Table 2.2:* The proportion of participants in each conditions that reported each strategy.

| | Strategies | | | | | |
|---|---|---|---|---|---|---|
| | CJ | 2D | UD | II/OS | Other | None |
| Overall | | | | | | |
| Rule-based | 0.325 | 0.1 | 0.15 | 0 | 0.15 | 0.225 |
| Information-integration | 0.3 | 0.15 | 0.3 | 0 | 0.025 | 0.125 |
| By condition | | | | | | |
| Rule-based | | | | | | |
| Feedback | 0.45 | 0.05 | 0.1 | 0 | 0.1 | 0.2 |
| Observation | 0.2 | 0.15 | 0.2 | 0 | 0.2 | 0.25 |
| Information-integration | | | | | | |
| Feedback | 0.4 | 0.25 | 0.25 | 0 | 0.15 | 0 |
| Observation | 0.2 | 0.15 | 0.35 | 0 | 0.05 | 0.25 |

Strategies: CJ=Conjunction, 2D=Generic two-dimensional strategy, UD=Unidimensional strategies, II/OS=Information-integration or overall similarity.

people who reported strategies) was significantly different from those in the information-integration conditions who identified the correct strategy (0% of the people who reported strategies), $\chi^2(1) = 15.20$, $p < .001$.

In sum, although participants found neither category structure trivial to verbalise, participants in the rule-based category structure conditions were more able to verbalise the underlying category structure than those in the information-integration conditions. Thus, these analyses largely support the assertion that the rule-based category structure is more readily verbalisable than the information-integration category structure.

**Comparison of model-based strategies with verbal reports**

The model-based analyses and verbal reports used here are complementary approaches that both aim to determine how participants are completing the task. However, from the summaries of these analyses above, it appears that they are partially inconsistent with each other. To examine the degree of correspondence between these approaches, I compared the strategy each participant was assigned using the model-based analysis with the one they reported using after the experiment (Table 2.3).

As can be seen, for the rule-based strategies (unidimensional, two-dimensional and conjunction) the model-based analyses and verbal reports match reasonably well. This is not the case for the diagonal (GLC) strategy and participants' reports of implicit or overall

*Table 2.3:* Comparison of the models assigned to each participant in the model-based strategy analysis with those they reported using in Experiment 1.

| Model-based strategy | Verbal strategy reports | | | |
|---|---|---|---|---|
| | CJ | UD | 2D | II/OS |
| Condition | | | | |
| Rule-based | | | | |
| CJ | 10 | 4 | 0 | 0 |
| UD | 1 | 6 | 3 | 0 |
| GLC | 2 | 2 | 3 | 0 |
| Information-integration | | | | |
| CJ | 5 | 4 | 0 | 0 |
| UD | 1 | 8 | 5 | 0 |
| GLC | 6 | 2 | 4 | 0 |

Strategies: CJ=Conjunction, UD=Unidimensional, GLC=General linear classifier. Verbal report strategies: CJ=Conjunction, UD=Unidimensional, 2D=Generic two-dimensional strategy, II/OS=Information-integration or overall similarity.

similarity responding; all participants that were assigned to the diagonal (GLC) strategy in the model-based analysis reported using an explicit rule-based strategy. One possible explanation for this disparity is that participants were using an implicit, diagonal (GLC) based, strategy but were unable to describe it correctly. Although, this may be unsurprising given that it is implicit, it seems unlikely given that in previous, different but related, work participants were able to report this type of strategy (Wills et al., 2013). Alternatively, it may be that the diagonal (GLC) strategy is more inclusive than the other models, and so results in participants that are using a rule-based strategy being assigned to the diagonal (GLC) merely because they could not be assigned to another type of strategy. This later hypothesis is supported by the Akaike weight for the diagonal (GLC); this model wins by a much lower margin than the others (see model-based analysis section above). This hypothesis is examined in detail in Chapter 5.

### 2.2.4 Discussion

As predicted by COVIS, Ashby et al. (2002) reported that performance with feedback training was superior to observational training when learning an information-integration category structure, whereas for a unidimensional rule-based category they found that these training types resulted in comparable performance. In contrast, in Experiment 1 I found that learning performance was better with feedback training than observational training to a similar degree for both category structures. The Bayesian Analysis veri-

fies that there is truly no difference in learning performance between the two category structures, rather than just a lack of statistical power (Dienes, 2011). This pattern of performance is not consistent with the claim that there are two systems of category learning that are differentially affected by training type. The state-trace analysis shown in Figure 2.3 also does not provide any evidence for a dual-system approach. It consists of a single, monotonically increasing curve, which is interpreted as evidence that performance in this experiment can be described by a single system of category learning.

COVIS could encompass the pattern of performance found in Experiment 1 if participants resorted to using the Procedural System for both category structures. However, this hypothesis is not supported by the verbal report analysis as this found that participants were equally likely to be able to report a strategy in all conditions, but that fewer participants were able to describe the optimal strategy in the information-integration conditions than in the rule-based conditions. Similarly, the model-based analysis indicates that the conjunction model best fits more participants in the rule-based condition than the information-integration condition, and a diagonal (GLC) model best fits more participants in the information-integration condition than the rule-based condition. Therefore, the results of Experiment 1 appear inconsistent with COVIS.

## 2.3 Experiment 2

### 2.3.1 Introduction

Ashby et al. (2002) found an interaction between training type and category structure. They argued that this pattern of results supported COVIS. However, Ashby et al. included several confounds in their design that complicate interpretation of their results: the number of stimulus dimensions relevant to categorisation, category separation and error rates. When these were controlled for in Experiment 1, feedback training was superior to observational training when learning both rule-based and information-integration categories–a pattern of results not consistent with COVIS. The key difference between Experiment 1 and Ashby et al.'s findings is the appearance of a feedback training advantage for rule-based categories. In Experiment 2, I aimed to determine which of the controlled for confounds might have resulted in the differential effect of training type reported by Ashby et al..

The number of dimensions relevant to classification seemed to be the most likely cause of this difference. This is because Ashby et al. (2002) manipulated category separation

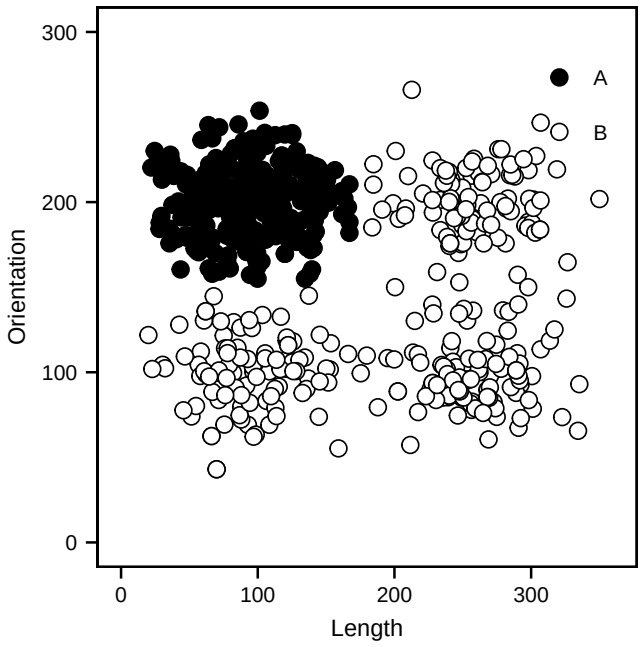


*Figure 2.4:* The conjunction category structure used in Experiment 2 in abstract stimulus space. Each point represents a particular stimulus. Filled circles represent Category A and unfilled circles represent Category B.

and error rates in a second experiment and still did not find a statistically significant difference in performance due to training type. Therefore, to discriminate dimensionality from the other factors, the number of relevant dimensions in the category structure were maintained whilst category separation and error rates were varied (see Figure 2.4). Category separation was increased. Error rates were reduced by scaling the length dimension to increase perceptual discriminability along that dimension. Additionally, on each trial the stimulus, category label and inter-trial interval were increased to 1000ms as a pilot study indicated that simply changing the perceptual discriminability did not improve average errors sufficiently.

If increased error rates or reduced category separation are the cause of the difference in learning rule-based categories between Experiment 1 and Ashby et al. (2002) then the difference between training type should disappear in this experiment. However, if the locus of the difference is the number of relevant dimensions for optimum classification then the advantage for feedback training over observational training should remain.

### 2.3.2 Method

#### Participants

40 participants (10 male) were recruited from the Plymouth University participation pool and were paid £8 for their participation.

#### Design

The experiment had 2 between-subjects conditions (training type: feedback, observational). 20 participants were randomly assigned to each condition. Participants in the feedback condition were first shown the stimulus, then were given an opportunity to respond, followed finally by the correct category label. Participants in the observational condition were first shown the correct category label, then the stimulus, followed finally by the opportunity to respond. Category learning was measured by the proportion of correct responses in each test block.

#### Stimuli

This version of the experiment utilised a conjunction category structure similar to that used in Experiment 1. However, the category structure was altered to make learning easier (Figure 2.4). To generate the category structure, four sets of points, 300 from the Category A distribution and 100 each from the other three, were randomly selected from bivariate normal distributions defined using the parameters listed in Table 2.4. Any points that were over 2.25 standard deviations away from the mean of the distribution in the direction of the category boundary were resampled. Then, as Experiment 1 indicated that the orientation of the line stimuli appeared more salient than line length to participants, the distribution was scaled so that the lines varied between 20 and 350 points in arbitrary units. This resulted in the distribution of points found in Figure 2.4.

*Table 2.4:* Parameters used to generate the initial stimulus distribution for Experiment 2.

| Category | Parameters | | | |
|---|---|---|---|---|
| | $\mu_l$ | $\mu_o$ | $\sigma_l$ | $\sigma_o$ |
| A | 100 | 200 | 20 | 20 |
| B | 100 | 100 | 20 | 20 |
| B | 200 | 100 | 20 | 20 |
| B | 200 | 200 | 20 | 20 |

Note: $\mu_l$=Mean length, $\mu_o$=Mean orientation, $\sigma_l$=Standard deviation of length, $\sigma_o$=Standard deviation of orientation dimension.

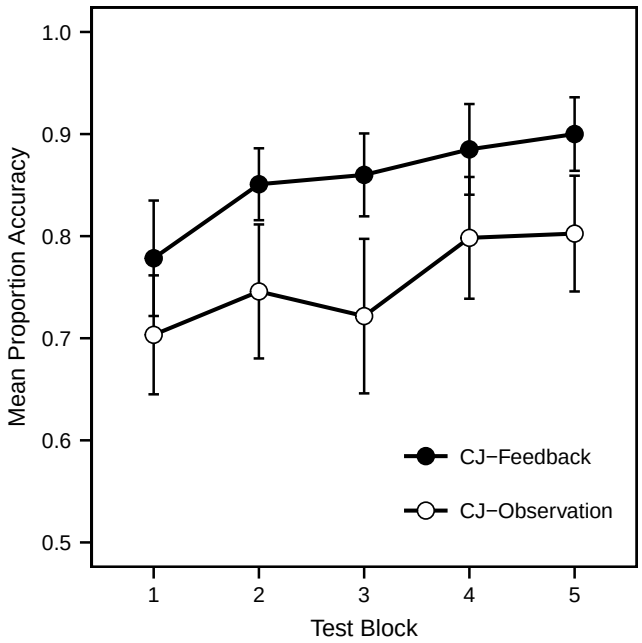


*Figure 2.5:* Average accuracy for each test block in each condition in Experiment 2. Error bars are 95% difference-adjusted confidence intervals (Baguley, 2012).

**Procedure**

There were two additional changes between Experiment 1 and Experiment 2. First, the length of each section of trial was increased: the intertrial interval, stimulus presentation and category label presentation were all increased from 500ms to 1000ms.

Second, the verbal report questionnaire was refined based on our experience of coding the Experiment 1 responses in order to elicit clearer descriptions. Here, participants were asked to "Imagine that another participant was asked to complete the experiment exactly as you did. What instructions would you give them so that they could exactly copy your pattern of responding? Please try to be as precise as possible."

### 2.3.3 Results

**Null-hypothesis significance testing**

Following Ashby et al. (2002) and Experiment 1, analyses were conducted on the final test block of the data from all participants. Conducting the analyses across all test blocks lead to the same conclusions. No participant failed to reach 50% accuracy by the final test block. Figure 2.5 shows mean proportion accuracy in each condition for all test blocks. The trial-level raw data for Experiment 2 are archived at `www.willslab.co.uk/ply27`.

There was a statistically significant main effect of training type $F(1,38) = 4.83$, $\eta^2 = 0.09$,

*Table 2.5:* The proportion of participants in each condition fit by each GRT model.

| Condition | Strategies | | | | |
|---|---|---|---|---|---|
| | CJ | UDO | UDL | GLC | RND |
| Feedback | 0.9 | 0 | 0 | 0.05 | 0.05 |
| Observation | 0.65 | 0 | 0.1 | 0.05 | 0.2 |

Strategies: CJ=Conjunction, UDO=Unidimensional strategy based on stimulus orientation, UDL=Unidimensional strategy based on stimulus length, GLC=General linear classifier, RND=Random.

$p = 0.034$. Participants trained using feedback had higher performance, $M = 0.85$, $SD = 0.13$, than those given observational training $M = 0.75$, $SD = 0.19$.

**Model-based strategy analysis**

The proportions of participants using each model in each condition are in Table 2.5. From this we can see that the majority of participants in both conditions have been identified by the analysis as using either the correct conjunction strategy or another rule-based one. This supports the hypothesis that the participants are using an explicit, rule-based strategy. However, the proportions of participants in each condition that were assigned to the correct conjunction strategy are statistically different, $\chi^2(1) = 5.63$, $p = .018$. This indicates that participants were more successful at determining the underlying category structure in the feedback training condition than in the observational training condition.

Additionally, I looked at the performance of the best-fitting model relative to the performance of the competing models in terms of the Akaike weights. For the participants in the feedback training condition the mean normalized probability of using a rule-based strategy compared to information-integration or random strategies is 0.931, whereas for the observational training conditions the normalized probability is 0.770. This demonstrates that, as expected, participants are most likely to use rule-based strategies in these rule-based conditions and that these strategies were clear winners.

**Verbal report analysis**

The verbal reports were independently coded by the author (CERE) and an independent rater (ABI). Any discrepancies that were not due to human error were easily resolved through discussion.

Inter-rater reliability for judging whether or not each participant reported a strategy was high, $\kappa = .844$, $p < .001$, whereas judgements as to which strategy they were reporting

*Table 2.6:* The proportion of participants in each condition that reported each strategy.

| | Strategies | | | | | |
|---|---|---|---|---|---|---|
| | CJ | 2D | UD | II/OS | Other | None |
| Overall | | | | | | |
| Rule-based | 0.625 | 0.20 | 0.10 | 0 | 0 | 0.075 |
| Rule-based | | | | | | |
| Feedback | 0.70 | 0.20 | 0.05 | 0 | 0 | 0.05 |
| Observation | 0.55 | 0.20 | 0.15 | 0 | 0 | 0.10 |

Strategies: CJ=Conjunction, 2D=Generic two-dimensional strategy, UD=Unidimensional strategies, II/OS=Information-integration or overall similarity.

were reasonable, $\kappa = .595$, $p < .001$. The majority of discrepancies appeared to be due to different interpretations of how participants might be expected to describe a conjunction category structure. The coded strategies described by participants are shown in Table 2.6.

There were no significant differences between conditions in the number of participants who did not report a strategy, $\chi^2(1) = 0.36$, $p = .548$. There was also no significant difference between condition in those who reported the correct conjunction category, $\chi^2(1) = 0.96$, $p = .327$. Therefore, participants in both conditions were capable of not only coming up with a strategy, but the majority were also able to correctly describe the category structure.

**Comparing model-based analyses with verbal reports**

As in Experiment 1, I also looked at the degree of correspondence between the verbal reports given by participants and the model that best fit their responses as determined by the model-based analysis (Table 2.7).

In this experiment, the verbal reports matched the model-based analysis reasonably well; the majority of participants that reported using a conjunction strategy were also assigned to this in the model-based analysis. Furthermore, as might be expected in learning a rule-based category structure, no participants reported using implicit or overall similarity responding or were best described, in the model-based analysis, by the diagonal (GLC) model.

*Table 2.7:* Comparison of the models assigned to each participant in the model-based analysis with those that they reported using.

| Model-based strategy | Verbal strategy reports | | | |
|---|---|---|---|---|
| | CJ | UD | 2D | II/OS |
| CJ | 22 | 2 | 6 | 0 |
| UD | 0 | 1 | 2 | 0 |
| GLC | 0 | 0 | 0 | 0 |
| RND | 2 | 2 | 0 | 0 |

Strategies: CJ=Conjunction, UD=Unidimensional, GLC=General linear classifier, RND=Random. Verbal report strategies: CJ=Conjunction, UD=Unidimensional, 2D=Generic two-dimensional strategy, II/OS=Information-integration or overall similarity.

### 2.3.4 Discussion

The key difference between Experiment 1 and Ashby et al. (2002) was the appearance of an advantage for feedback training over observational training when learning a rule-based category structure. In Experiment 2, I aimed to determine which of the factors that varied between these two experiments was responsible for this difference. To do this, Experiment 2 compared performance with feedback and observational training when learning a two-dimensional category, with reduced error rates and increased category separation compared with the category structure used in Experiment 1. Under these conditions, the advantage of feedback training over observational training remained. In addition, the model-based and verbal reports indicate that the majority of participants in both conditions were able to use and verbally describe a conjunction strategy. This indicates that the interaction between training type and category structure found by Ashby et al. appears to be due to differences in dimensionality between the category structures.

## 2.4 General discussion

Ashby et al. (2002) reported that feedback training was superior to observational training for an information-integration category structure, but that the two training types were comparable for a rule-based category structure. This dissociation has widely been taken as support for the COVIS dual-process theory of category learning (Ashby et al., 1998, 2011; Ashby & Valentin, 2016) and is the most cited, un-critiqued behavioural support for this model. According to the COVIS framework, the critical manipulation in Ashby et al. (2002) is that rule-based category structures are easily verbalisable, while information-integration categories are not and that this results in participants learning these two types

of category using different category learning systems. These two systems incorporate feedback differently, therefore accounting for the Ashby et al. findings. However, there were several non-essential differences between the category structures used by Ashby et al., which casts doubts on whether verbalisability is the key factor in eliciting a differential effect of training type on learning performance.

In Experiment 1, I successfully maintained the between category structure difference in verbalisability while matching them for (a) the number of relevant stimulus dimensions, (b) category separation, and (c) overall performance. I did this by combining the procedures of Ashby et al. with two-dimensional category structures adopted from more recent work in the COVIS framework (specifically Filoteo et al., 2010). Once these extraneous factors were controlled for, the category structure by training type interaction found by Ashby et al. did not appear: learning of both category structures was better with feedback training than observational training. Experiment 2 also found a training type difference in learning the two-dimensional rule-based structure when this structure was broadly matched, in terms of category overlap and overall performance, with the rule-based structures used by Ashby et al.. This indicates that the appearance of a differential effect of training type on rule-based learning in these experiments appears to be due to the two-dimensional nature of the conjunction structure. These experiments demonstrated an advantage for feedback training over observational training for not only information-integration categories, but also for two-dimensional rule-based categories. This has considerable import if you consider that a substantial amount of classroom instruction relies on observational training; these results indicate that instructors should aim to change their style when teaching difficult concepts.

### 2.4.1 Alternative explanations

This result also has implications for the COVIS theory of category learning because the results of the current experiments are not predicted by COVIS. In Experiment 1, COVIS predicts a greater feedback advantage for the information-integration structure than the rule-based structure, but both conditions benefit from feedback training to a similar degree. In Experiment 2, COVIS does not predict a feedback advantage, yet one is observed. How, then, might the results of both Ashby et al. (2002) and the current paper be explained?

The first thing to explain is why feedback training is superior to observational training.

Any theory that presumes learning is driven by prediction error (see e.g. Wills et al., 2009, for a review) should be able to accommodate this result because, in observational training, there is nothing to predict. The ALCOVE model (Kruschke, 1992) is one of several possible category learning models in which learning is driven by prediction error, as is the striatal pattern classifier (Ashby & Waldron, 1999) that forms the basis of Ashby's explanation of why a feedback advantage is sometimes observed.

Second, I need to explain why a benefit of feedback training is sometimes not observed. One possibility is that such findings represent absence of evidence rather than evidence of absence. In Ashby et al.'s first experiment, performance on the harder, observational, training condition is close to ceiling, potentially obscuring the effect. In addition, Ashby et al. report a significant feedback advantage for the unidimensional category structure in the first test block (Ashby et al., 2002, p. 673), which smoothly reduces throughout training until it disappears in the final block (Ashby et al., 2002, Figure 3). Ashby et al.'s conclusions are based on the final block. In Ashby's second experiment, there is a numerical trend in the direction I would predict, sample sizes are small, and only one of the two counterbalance conditions were below ceiling. Thus, one possibility is that feedback is always advantageous in rule-based category learning, but that some experiments fail to reveal this due to methodological issues.

Another possibility is that the feedback advantage is genuinely absent for single-dimension rule-based category structures, or at least much smaller than it is for multi-dimensional category structures (rule-based or otherwise). Although further research would be required to make this claim securely, it is interesting to speculate how such an effect might be explained if it were to be confirmed. One possibility is that the size of the feedback advantage is related to how effortful the classification is (Markant, Ruggeri, Gureckis, & Xu, 2016). Dimensional Summation theory (Milton & Wills, 2004) predicts that single-dimension classification is less effortful than multi-dimensional classification, and this prediction has been supported in multiple studies (e.g. Milton et al., 2008; Wills et al., 2013, 2015).

In summary, COVIS predicts that there should be an interaction between training type and category structure, with a smaller difference between training types when learning a readily verbalisable category structure compared to one that is difficult to verbalise. However, the available evidence (from both Ashby et al. and the current studies) indicates

that the pattern of performance on these tasks might be better explained by an interaction of training type and the number of dimensions relevant to classification. Of course, these experiments have not completely disentangled verbalisability from dimensionality. In order to do this, one would have to examine the effect of training type on a unidimensional, difficult to verbalise category. This would be difficult as it is hard to conceive of a unidimensional category structure that would be hard to verbalise without redefining what is meant by a stimulus dimension.

More generally, although there is reasonable support for the idea that providing an opportunity for error improves learning (Grimaldi & Karpicke, 2012; Kornell, Hays, & Bjork, 2009; Potts & Shanks, 2014), such an effect is not always seen even in multidimensional category structures (Newell et al., 2007) and, in some memory tasks, the effect is even reversed (Haslam, Hodder, & Yates, 2011). Neither COVIS, nor my alternative explanation, fully captures these results. Further empirical work is required to clearly identify the conditions under which feedback training is superior to observational training.

### 2.4.2 Dimensionality

As discussed above, it seems likely that it is the problem dimensionality, rather than the problem verbalisability, that drives the results of Ashby et al. (2002) and the current paper. The comparison of a unidimensional rule-based category structure with a 45-degree rotation of that structure in stimulus space has formed the basis of a large number of experiments by Ashby and colleagues. The comparison is initially appealing, because the two structures are in various formal senses identical (e.g. an optimal classifier performs equally well on both structures), yet one is easy to verbalise while the other is hard to verbalise. However, the two structures are not matched on the number of psychological stimulus dimensions relevant to the classification. This raises the broader question of whether a failure to control problem dimensionality underlies other apparently COVIS-supporting dissociations.

A proponent of COVIS might suggest that dimensionality is unlikely to be driving the difference of these results and those of Ashby et al. (2002) on the basis that pigeons find the two problems equally difficult (Smith et al., 2011), the implication being that if a non-verbal species finds these two problems equally difficult then it must be the verbalisability of the problems rather than their dimensionality that is important. However, even in non-verbal species, a necessary condition of a unidimensional problem being easier than a

two-dimensional problem is that the stimulus dimensions are psychologically separable. Without separability, there is no meaningful psychological sense in which the two problems differ in dimensionality. Smith et al. (2011) provide no compelling evidence that their stimuli are separable for pigeons.

Another possible response to my claim that dimensionality is the critical factor is to point out that many of the more recent COVIS-supporting dissociations make use of a two-dimensional rule-based structure, thus equating problem dimensionality between rule-based and information-integration problems (e.g. Maddox, Bohil, & Ing, 2004; Maddox, Filoteo, Hejl, & Ing, 2004; Maddox & Ing, 2005; Maddox, Filoteo, & Lauritzen, 2007; Maddox et al., 2008; Zeithamova & Maddox, 2006) and dissociations, predicted by COVIS, still emerge. However, this evidence is not, perhaps, as compelling as it first appears and in recent years it has attracted substantive critiques on a variety of bases from separate labs (e.g. Dunn et al., 2012; Newell et al., 2010, 2013; Stanton & Nosofsky, 2013; Zaki & Kleinschmidt, 2014). This explanation is, therefore, entirely compatible with the existing evidence.

### 2.4.3 Conclusions

In summary, the current chapter casts doubt on the interpretation of the dissociation found by Ashby et al. (2002). The current experiments demonstrated an advantage for feedback training over observational training not only for information- integration categories, but also for two-dimensional rule-based categories. Therefore, category structure dimensionality, rather than verbalisability, appears to be the key factor driving the appearance of an interaction between category structure and training type in the original study. This paper, therefore, adds to the growing literature (e.g., Dunn et al., 2012; Newell et al., 2010, 2013; Stanton & Nosofsky, 2007, 2013) that casts doubt on the validity or interpretation of the experimental evidence for the COVIS model of category learning.

# Chapter 3

# Training order

In Chapter 2, I reported two experiments that examined a paper that was argued to support the dual-system of category learning COVIS. These experiments re-examined work by Ashby et al. (2002) that had found a differential effect of training type on rule-based and information-integration category learning. Contrary to Ashby et al., I found evidence that feedback training is superior to observational training for both category structures, which is more consistent with a single-system account.

As well as the experimental findings, these experiments contained other evidence pertinent to the validity of the COVIS model. The experiments in Chapter 2 also examined the strategies that participants used to complete the task. When looking at the model-based analysis informed by GRT (Ashby & Townsend, 1986; Ashby & Gott, 1988), a proportion of participants in the information-integration category were found to be using rule-based strategies. Additionally, the verbal reports that participants gave all described using rule-based approaches to the task. This includes the participants that had been found by the model-based approach to be using the diagonal (GLC) strategy, which is hypothesised to represent implicit learning within the COVIS literature.

These strategy findings raise an interesting question: do participants ever switch from the Verbal System to the Procedural System? It may be possible that the Procedural System does exist, but the experiments reported in Chapter 2 and Ashby et al. (2002) just failed to encourage participants to switch to it. To examine this hypothesis more closely, in this chapter I examine an experiment, purported to support the COVIS model, that relies on the timing of participants switching between systems to explain its findings (Spiering & Ashby, 2008).

## 3.1 Introduction

Spiering and Ashby (2008) found that the effect of changing the order of the exemplars seen in training depends on the type of category structure being learned. In two exper-

iments, participants learned either an information-integration or a conjunction category structure with one of three training orders: easy-to-hard, hard-to-easy or random. In the easy-to-hard conditions, participants first saw the easy stimuli that were furthest from the category boundary, followed by the intermediate ones, and then the hard stimuli that were closest to the category boundary. In the hard-to-easy conditions, participants saw the stimulus types in the opposite order. In the random conditions, participants saw all the stimuli in every training block. Finally, all the participants were tested on all the stimuli in the category structure. Spiering and Ashby found that when learning the information-integration category structure participants who were given hard-to-easy training had better final performance than those given easy-to-hard or random training. However, for participants who learned the rule-based conjunction category structure they found no statistically significant differences between training order conditions.

Spiering and Ashby (2008) argued that the differential effect of training order on the learning of these two category structures is predicted by COVIS.* Recall from Chapter 1 that COVIS assumes that the Verbal System mediates optimal performance in verbalisable rule-based category structures, whereas the Procedural System mediates optimal performance in difficult-to-verbalise category structures such as an information-integration structure. According to COVIS, all participants begin learning using simple verbalisable rules mediated by the Verbal System, only switching to the Procedural System if these rules result in poor performance. For this category structure, a simple unidimensional rule-based strategy would result in excellent performance with the easy stimuli (100%), but would result in poor performance when classifying the moderate (80%) and difficult (60%) stimuli closer to the category boundary. Therefore, participants in the easy-to-hard condition might delay switching to the optimum Procedural System because using the Verbal System initially resulted in high levels of accuracy. Whereas, the participants in the hard-to-easy condition would score lower using a simple unidimensional rule and so would more quickly realise that "no explicit strategies will succeed" (p. 1171). They would then more quickly switch to using the Procedural System to learn the structure and so score more highly. For the conjunction category structure, COVIS predicts no effect of training order because all participants are already using the Verbal System at the beginning of training, which is the optimal system for that category structure. Spiering and Ashby conclude their results indicate that "if the optimal rule is not easily verbalised . . . the

*Note that the random condition is not relevant for discussions about the validity of COVIS and so is not included in the following.

most effective training procedure might be to begin with difficult examples" (p.1176).

However, Spiering and Ashby's (2008) findings are also consistent with two simpler, single-system theoretical accounts. The first account hypothesises a rule-based mechanism similar to the Verbal System of COVIS. However, this rule-based mechanism can implement rules more complex than those hypothesised to be implemented by the Verbal System. Spiering and Ashby state that "no explicit strategies will succeed" (p. 1171), however, the validity of this statement depends on which set of strategies they are referring to. While it is certainly true that no unidimensional strategy could score 100% on the information-integration task in Spiering and Ashby, some more complex rule-based strategies can score highly. For example, the highest accuracy a conjunction strategy could achieve is 96.7% which, although not perfect, is still extremely high. Other strategies not included in the COVIS literature, such as rule-plus-exception strategies (e.g. Nosofsky, Palmeri, & McKinley, 1994), can also perform well.

So how might using complex rule-based strategies predict the pattern of results found by Spiering and Ashby (2008)? The argument is very similar to that used in the original paper; however rather than switching between learning systems, here participants switch between strategies. For the information-integration category structure, training participants on the easy stimuli teaches them a simple rule that generalises to the other stimuli poorly. In contrast, training participants on the hard stimuli teaches them a more complex strategy that generalises to the other stimuli well. This results in higher performance for participants in the hard-to-easy condition than in the easy-to-hard condition. On the other hand, for the rule-based conjunction structure, good performance on both the easy and hard stimuli requires the participants to use the same conjunction strategy (there are no sub-optimum strategies that are reasonably successful). Therefore, there are no differences between training type conditions.

The other alternative single-system possibility cannot be ruled out due to a confound in Spiering and Ashby's (2008) experimental design (see Figure 3.1 for their accuracy scores across training). Spiering and Ashby argue that the effect of training order on information-integration category learning is due to the type of stimuli the participant saw *initially* (in Block 1). However, the stimuli in the training block just prior to the final test block (Block 3) are hard in the easy-to-hard condition and easy in the hard-to-easy condition. Therefore, it is possible that participants' performance in the test phase following

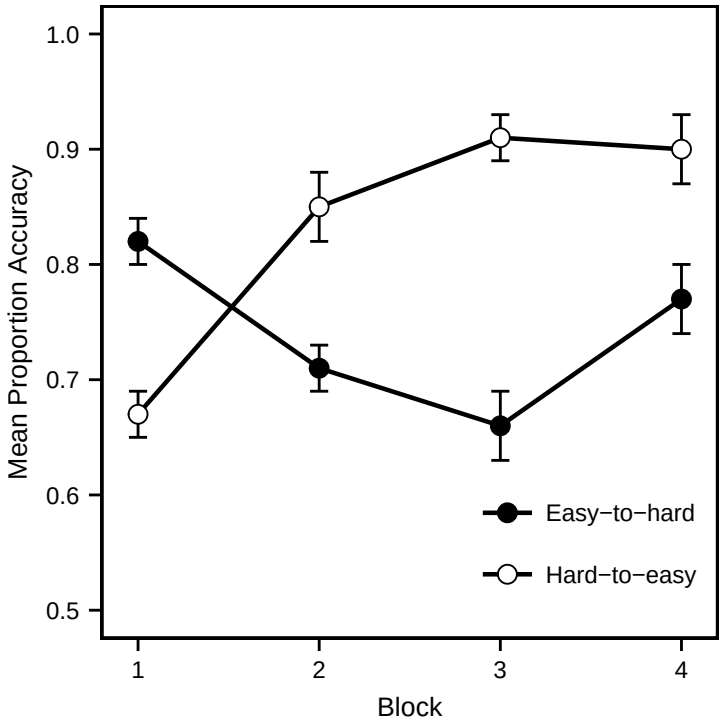


*Figure 3.1:* Estimated accuracy scores for the two experimental conditions from Experiment 1 of Spiering and Ashby (2008)

training reflects the most *recent* rather than the initial training. In other words, there may be a confound between primacy and recency. If this were the case, their results would actually be similar to an opposite effect: Transfer Along a Continuum (TAC, also known as easy-to-hard transfer; Suret & McLaren, 2003). Here, initial training on an easy version of a discrimination (compared to training on a hard version) results in superior performance on the hard discrimination at test (Lawrence, 1952). In humans, TAC has been found with a large variety of stimuli, for example, faces (Suret & McLaren, 2003), mammograms (Hornsby & Love, 2014) and birdsong (Church, Mercado, Wisniewski, & Liu, 2013).

In summary, the second alternative explanation of the Spiering and Ashby (2008) results is that performance is based on the training block just prior to test rather than the first training block. This means that the apparent effects of initial easy training might actually be the effects of recent hard training and vice versa. This explanation is only partially consistent with the data as Spiering and Ashby found the same effect at Block 2 (where both groups were learning the moderately difficult stimuli) as they did at Block 4. Indeed, the effect in Block 2 is more valid than the effect in Block 4 as there is no confound between primacy and recency. Therefore, it is unlikely that the effect Spiering and Ashby reported in Block 4 is due to the most recent training block (Block 3). Nonetheless, this

confound makes interpretation of their results difficult. To avoid the same interpretative difficulty the first three experiments reported below have only two blocks. This ensures that participants' performance can only be attributed to the training they initially received.

## 3.2 Experiment 3

In Experiment 3, I aim to replicate the first two blocks of Experiment 1 in Spiering and Ashby (2008). Spiering and Ashby looked at the effect of training order on learning both an information-integration category structure (Experiment 1) and a conjunction category structure (Experiment 2). I chose to focus on the information-integration structure as Spiering and Ashby found evidence of a counterintuitive learning effect of considerable theoretical interest with this structure (they found no effects of training order with a conjunction category structure).

Both a rule-based single-system approach and COVIS would predict that participants in the hard-to-moderate condition would end up with better final performance than those in the easy-to-moderate condition. However, a more intuitive account, which is consistent with the TAC literature, would predict that training on the easy stimuli would result in superior performance than training on the hard stimuli.

### 3.2.1 Method

#### Participants

The participants were 40 undergraduate psychology students recruited from Plymouth University participant pool and were randomly assigned to one of the two conditions (N=20 in each). They received 1 research credit in exchange for their participation.

#### Category structure and stimuli

The stimuli were sine-wave gratings displayed on a grey background that were identical to those used in Spiering and Ashby (2008). The stimuli used are shown in Figure 3.2.

#### Design

This experiment had a single between-subjects factor with two levels: easy-to-medium and hard-to-medium. In Block 1, participants in the easy-to-moderate condition were shown stimuli that were easy to classify as the stimuli were furthest from the decision boundary. In contrast, participants in the hard-to-moderate condition were shown stimuli that were difficult to classify as the stimuli were closest to the decision boundary. Then

in Block 2, participants in both conditions were shown stimuli that were moderately difficult to classify. The design was identical to Spiering and Ashby (2008), however, unlike Spiering and Ashby we did not include a random condition as this was not relevant to our theoretical predictions.

**Materials**

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

**Procedure**

On each trial, a stimulus and the category labels were displayed on a white background until the participant had responded by pressing either the 'D' or 'L' key. If a participant failed to respond after 5000ms had passed, a screen displaying 'PLEASE RESPOND FASTER' was shown to them. If the participant responded, 500ms of audio feedback was played to them over headphones. For correct responses, the tone was a 262Hz sine-wave, which sounds similar to a tuning fork. For incorrect responses, the tone was a 400Hz saw-tooth, which sounds harsher than a pure tone. The inter-trial interval was 1500ms.

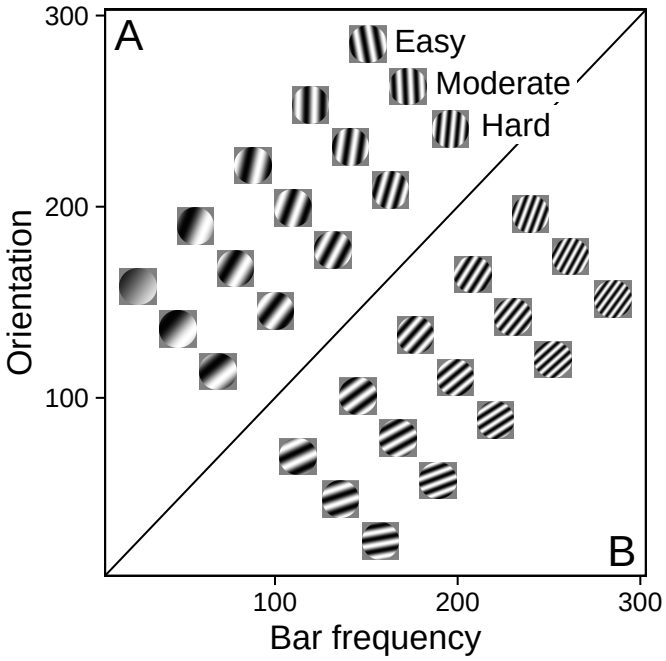


*Figure 3.2:* The stimuli used in Experiment 3 represented in abstract stimulus space. The diagonal line represents the optimal category bound. The two categories are labelled A and B. Also included are the stimulus difficulties for Category A: the stimuli furthest from the decision boundary are the easy stimuli and those closest to the category boundary are the hard stimuli.

There were 2 blocks of 150 trials each, resulting in a total of 300 trials. In each block, 10 stimuli were presented in a random order 15 times. The stimuli presented in each block depended on the condition to which the participant was assigned. In Block 1, participants in the easy-to-medium condition were shown only the easy stimuli, those furthest from the decision boundary; participants in the hard-to-medium condition were shown only the hard stimuli, those closest to the decision boundary. In Block 2, the participants in both conditions were shown the moderately difficult stimuli in a random order, 15 times each, with feedback.

Additionally, after the experiment was completed, participants were asked to fill in a questionnaire. This aimed to determine which strategy they had used to categorise the stimuli. They were asked to "Imagine that another person was asked to complete the experiment as you did. What instructions would you give them so that they could exactly copy your pattern of responding?" They were given a large box in which to fill in their answer and asked to respond as precisely as possible.

**Analysis**

All trials for which the reaction time was greater than 5000ms were removed.

Throughout this chapter, as in Chapter 2, I also calculated Bayes Factors. This is because in null-hypothesis significance testing non-significant results are ambiguous: they could either be due to insufficient statistical power or due to the null hypothesis being correct (Dienes, 2011). It is important to be able to disambiguate between these two possibilities.

Recall that if the Bayes Factor is over three then the experiment has found evidence for the experimental hypothesis whereas if the Bayes Factor is less than a third, the experiment finds evidence for the null hypothesis (Jeffreys, 1961). A Bayes Factor of one indicates that the evidence is exactly neutral with respect to the experimental and null hypotheses (Dienes, 2011). Values between a third and three are typically interpreted as indicating that the experiment was not sensitive enough and no conclusions can be drawn.

The Bayes Factors for the accuracy data in the experiments in this chapter were calculated according to the procedure recommended by Dienes (2011) using the R script implemented by Baguley and Kaye (2010). The predicted differences between the easy-to-moderate and hard-to-moderate conditions were estimated from Spiering and Ashby (2008). For Block 1, I assumed a two-tailed normal distribution with a predicted mean

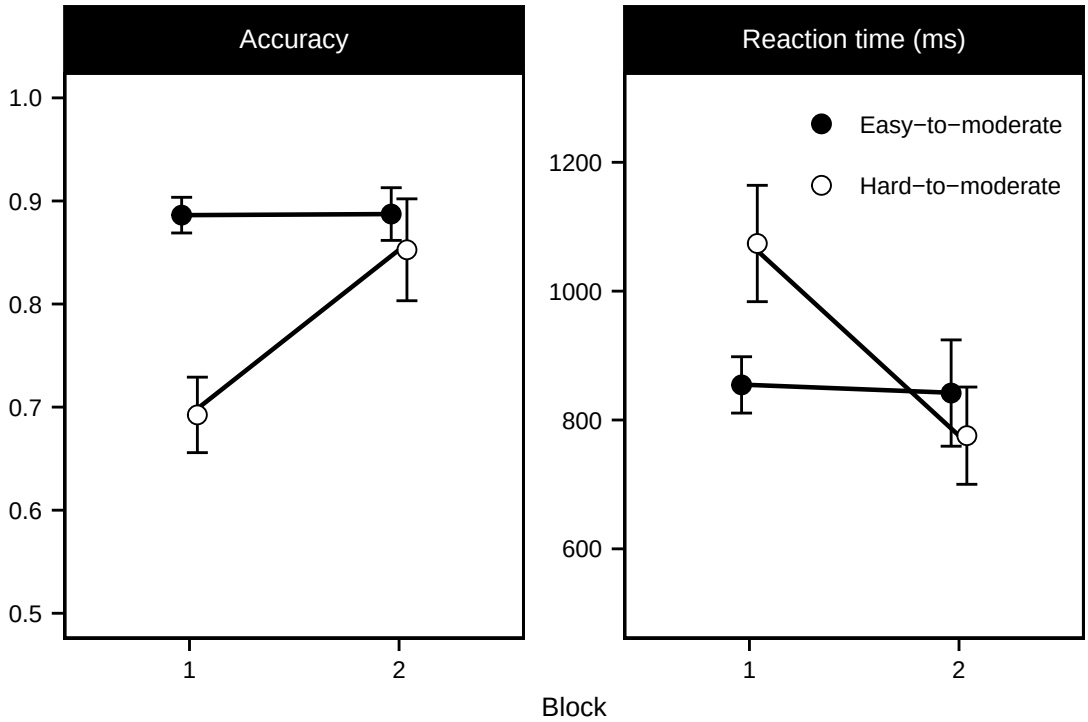


*Figure 3.3:* The average accuracy and reaction time for each block in each condition in Experiment 3. Error bars are difference-adjusted 95% confidence intervals (Baguley, 2012).

difference of 0.101 and predicted standard deviation of 0.054. For Block 2, I assumed a two-tailed normal distribution with predicted mean difference of -0.139 and standard deviation of 0.07.

Bayes Factors for the reaction time data were not calculated as average reaction times were not reported in Spiering and Ashby (2008).

### 3.2.2 Results

Three participants in the easy-to-moderate condition scored considerably below chance (less than 0.15 in Block 1 compared to chance at 0.50) and so were excluded from the following analyses. The results are displayed in Figure 3.3. The trial-level raw data for Experiment 3 will be available at `www.willslab.co.uk/ply22` once this experiment is published.

Bayes Factors were calculated for this experiment with the Block 1 sample mean difference set as 0.194, with a sample standard deviation of 0.027. For Block 2 the sample mean difference was 0.045, with a sample standard deviation of 0.037.

During Block 1, as expected, performance on the hard stimuli was indeed worse than performance on the easy stimuli, $t(35) = 6.71$, $d = 2.21$, $p < .001$, $BF = 1.75 \times 10^6$. How-

ever, this initial difference in accuracy had no effect on learning performance in Block 2, $t(35) = 0.88$, $d = 0.29$, $p = .387$. Indeed, there was substantial evidence for the null as the Bayes Factor was less than a third, $BF = 0.07$. The Bayes Factor remains in favour of the null hypothesis even if the predicted mean difference between the conditions is reduced by two thirds to $0.046$, $SD_{\text{diff}} = 0.023$.

The reaction time data was consistent with this. During Block 1, performance on the hard stimuli was slower than responding on the easy stimuli, $t(35) = 3.07$, $d = 1.01$, $p = .004$. However, by Block 2 this difference between conditions had disappeared, $t(35) = 0.88$, $d = 0.29$, $p = .384$.

Additional strategy analyses are reported in Section 3.6 below.

### 3.2.3 Discussion

Spiering and Ashby (2008) found that hard initial training improved performance on an information-integration categorisation task compared to easy initial training. They argued that this finding supported the COVIS model of category learning (Ashby et al., 1998, 2011). However, this finding was also consistent with an explicit rule-based approach that hypothesises that participants can use complex explicit rules. It was also partially consistent with the TAC effect due to a confound in Spiering and Ashby's experimental design: the effect they found in Block 4 could have been due to the stimuli the participants saw initially or those they saw most recently. In Experiment 3, I aimed to replicate the first two blocks of this experiment, thereby avoiding this experimental confound. In contrast to Spiering and Ashby, I failed to find an advantage for either training order. Indeed, I found Bayesian evidence for the null hypothesis. In other words, initially seeing the easy or hard stimuli had no effect on performance in the second block.

## 3.3 Experiment 4

It is possible that I failed to find an effect in Experiment 3 because several aspects of the procedure added additional noise, thereby obscuring the effect. First, it is possible that the unusual choice of feedback could have misled some of the participants. In both Spiering and Ashby and Experiment 3, the feedback was a 500ms tone administered over headphones. However, the mapping of tone pitch to the feedback was not intuitive: the higher tone indicated incorrect responses and the lower tone indicated correct responses. This is not common practice in other studies within the COVIS canon (e.g Ell & Ashby,

2006), more broadly in experimental psychology or even in other non-experimental settings such as game shows. So it is possible that this feedback may have been consistently misinterpreted by some participants. This idea is further supported by the fact that three participants had scores well below chance (<15%). This level of performance indicates that they learned the structure but pressed the wrong keys for each category. Furthermore, as Spiering and Ashby did not apply a learning criterion, it is possible that they also included participants like these in their analyses. The selective use of learning criteria has been a problem for COVIS evidence before (e.g. Newell et al., 2010).

Another feature of the procedure that may have added additional noise is the choice of stimuli. Newell et al. (2010) noted that Gabor patches "do not easily lend themselves to verbalization." This observation was supported by the verbal reports, which are discussed in more detail in Section 3.6. In Experiment 3, participants described several eclectic stimulus features that appeared to map onto the bar width dimension. These features included how "zoomed in" the stimulus was, whether the stimulus was symmetrical or not, how many bars there were and the amount of contrast between light and dark. Using these representations may undermine the inferences we wish to draw from this experiment. For example, it is possible that the "zoom" property maps onto the experimenter defined dimension of interest (bar width) in a non-linear way. This type of mapping corresponds to sub-regions of the stimulus space being stretched, which may alter the representation of the category structure. As there is no way of determining post-hoc how these novel dimensions map onto the experimenter defined dimensions, it is impossible to determine how these alternative representations of the stimuli may impact our experiment. To address these possibilities, in Experiment 4, I repeated Experiment 3 using line stimuli and visual feedback, i.e. 'Correct' or 'Incorrect!'

### 3.3.1 Method

#### Participants

The participants were 43 undergraduate psychology students recruited from the Plymouth University participation pool. They were randomly assigned to either the easy-to-moderate condition (N=20) or the hard-to-moderate condition (N=23). They received 1 research credit in exchange for their participation.

#### Category structure and stimuli

The abstract category structure was identical to that used in Experiment 1 of Spiering and Ashby (2008) and Experiment 3 above. However, this category structure was instantiated with black line stimuli that appeared on a white background. These stimuli varied in the length of the line and its orientation. The variation in the length of the lines were matched to the variation of line length in previous research (e.g. Edmunds et al., 2015; Filoteo et al., 2010). To do this, I calculated the linear scaling factor that would transform the bar frequency value to a corresponding line length value, such that the minimum and maximum values were 25 and 285 respectively. These values were the length of the line in pixels. The orientation of the lines was the same as the orientation of the sine-wave gratings in Experiment 3.

#### Procedure

The procedure for this experiment was identical to that of Experiment 3 aside from changing the feedback type. Rather than using 500ms tones, I displayed either 'Correct' or 'Incorrect!' in black in the centre of the screen for 500ms. Also, I did not give these participants the strategy questionnaire after training.

#### Analysis

All data analyses were conducted in R (R Core Team, 2015). All trials for which the reaction time was greater than 5000ms were removed.

### 3.3.2 Results

No participants scored below chance so the following analyses are conducted on all participants. The results are displayed in Figure 3.4. The trial-level raw data will be available at `www.willslab.co.uk/ply12` once this experiment is published.

For this experiment, the Bayes Factors were calculated using the same technique and prior as described in Experiment 3. Here, in Block 1 the sample mean difference was $0.181$, with a sample standard deviation of the difference of $0.023$. For Block 2, the sample mean difference was $0.025$, with a sample standard error of $0.021$.

During Block 1, as expected, performance on the hard stimuli was indeed worse than performance on the easy stimuli, $t(41) = 7.52$, $d = 2.30$, $p < .001$, $BF = 3.09 \times 10^7$. However, this initial difference in accuracy had no effect on learning performance in Block 2,

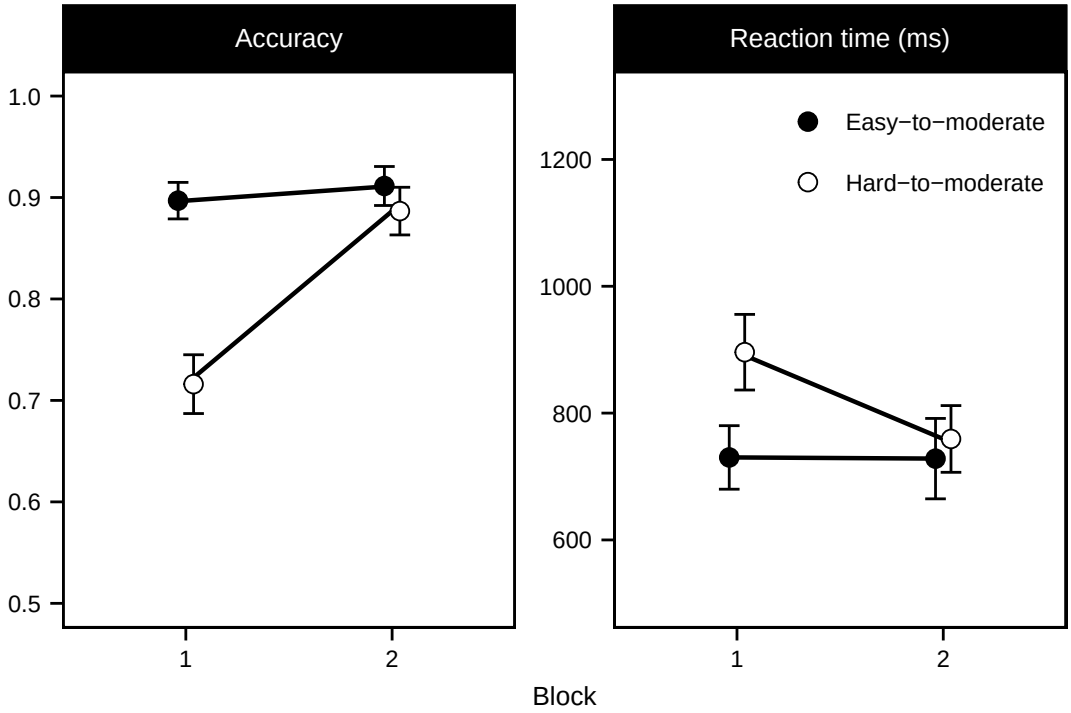


*Figure 3.4:* The average accuracy and reaction time for each block in each condition in Experiment 4. Error bars are difference-adjusted 95% confidence intervals (Baguley, 2012).

$t(41) = 1.18$, $d = 0.36$, $p = .877$. Indeed, there was substantial evidence for the null as the Bayes Factor was below a third, $BF = 0.05$. The Bayes Factor remains in favour of the null hypothesis even if the predicted mean difference is reduced by $6/7$ to $-0.02$, $SD_{\text{diff}} = 0.01$.

The reaction time data was consistent with this. During Block 1, performance on the hard stimuli was slower than responding on the easy stimuli, $t(41) = 2.96$, $d = 0.90$, $p = .003$. However, by Block 2 this difference between conditions had disappeared, $t(41) = 0.58$, $d = 0.18$, $p = .717$.

### 3.3.3 Discussion

Spiering and Ashby (2008) found that participants who were initially trained on a difficult discrimination had better final performance than participants who were trained on the easy version of that discrimination. In contrast, in Experiment 3 I failed to find a difference between training order conditions on final performance. I postulated that this may have been due to the unintuitive choice of feedback: the mapping from correct/incorrect to tone pitch was opposite to that usually seen in psychology experiments. In Experiment 4, I tested this hypothesis by changing the feedback from tones to visual feedback. However, this failed to make a difference: participants in both conditions still performed equally well

in the final block.

## 3.4 Experiment 5

In Experiments 3 and 4, I examined whether hard-to-medium, compared to easy-to-medium, training resulted in superior performance on an information-integration category structure. Contrary to Spiering and Ashby (2008), I failed to find an effect of initial training type on final performance: both easy and hard training resulted in the same level of performance in Block 2. Additionally, a Bayesian analysis found evidence for the null-hypothesis. This indicates that there is truly no difference between training order conditions. It also suggests that the effect reported by Spiering and Ashby may have been a false positive.

That being said, Spiering and Ashby (2008) found this difference at two points in their four block design: in Block 2 with only the moderately difficult stimuli and in Block 4 with all the stimuli. But, in Experiments 3 and 4, I only examined participants' performance on the moderately difficult stimuli. Therefore, it may be that if I include all the stimuli at test I may find an effect of training order on accuracy.

Additionally, including all the stimuli may correct for another discrepancy between my experiments and Spiering and Ashby (2008): performance levels. In Experiment 1 of Spiering and Ashby, Block 2 performance was lower than I observed in Block 2 of Experiments 3 and 4. Therefore, it may be that I failed to see an effect of initial training type because of a ceiling effect. By including all the stimuli, it was possible that it would increase the difficulty of the task as there would be more stimuli to learn. However, a pilot study still found performance higher than reported in Experiment 1 of Spiering and Ashby. Therefore, in Experiment 5 I also included a time pressure manipulation. Here, participants were only able to see the stimulus for 350ms before it was covered by a mask. I hoped this would drop Block 2 performance to levels comparable with Spiering and Ashby.

### 3.4.1 Method

#### Participants

The participants were 40 undergraduate psychology students recruited from the Plymouth University participation pool. They were randomly assigned to either the easy-to-all or hard-to-all (N=20 each) conditions. They were awarded 1 research credit in

exchange for their participation.

**Procedure**

The format of the experiment remained similar to that in Experiment 4. However, in this version, the participants were tested on all the stimuli in Block 2. Furthermore, the line stimulus was only shown for 350ms, after which a masked was displayed until the participant responded. The mask was constructed by placing every line stimulus three times on a white background with each end of the stimulus randomly displaced along both stimulus dimensions by a number of pixels randomly drawn from a uniform distribution between -120 and 120. This mask ensured that participants could only visually process the stimulus for the allotted 350ms.

As in Experiment 3, at the end of the experiment participants were also asked to report the strategy that they used to learn the category structure.

**Analysis**

All data analyses were conducted in R (R Core Team, 2015). All trials for which the reaction time was greater than 5000ms were removed.

### 3.4.2 Results

No participants scored below chance so the following analyses were conducted on all participants. The results are displayed in Figure 3.5. The trial-level raw data will be available at `www.willslab.co.uk/ply42` once this experiment is published.

For this experiment, the Bayes Factors were calculated using the same technique and prior as described in Experiment 3. Here, for Block 1, the sample mean difference was 0.227, with a sample standard error of 0.021. For Block 2, the sample mean difference was 0.013, with a sample standard error of 0.026. As the current experiment looked at performance on all stimuli (easy, moderate, and hard) in Block 2, some might argue that using a prior based on Block 4 of Experiment 1 of Spiering and Ashby (2008) might be more appropriate (as this was the block in their experiment in which all stimulus difficulties were presented). Use of this prior makes no difference to the conclusions drawn below.

During Block 1, as expected, performance on the hard stimuli was worse than performance on the easy stimuli, $t(36) = 10.40,\ d = 3.38,\ p < .001,\ BF = 1.13 \times 10^{10}$. However, this initial difference in accuracy had no effect on learning performance in Block 2, $t(36) = 0.48,\ d = 0.157,\ p = .684$. Indeed, there was substantial evidence for the null as

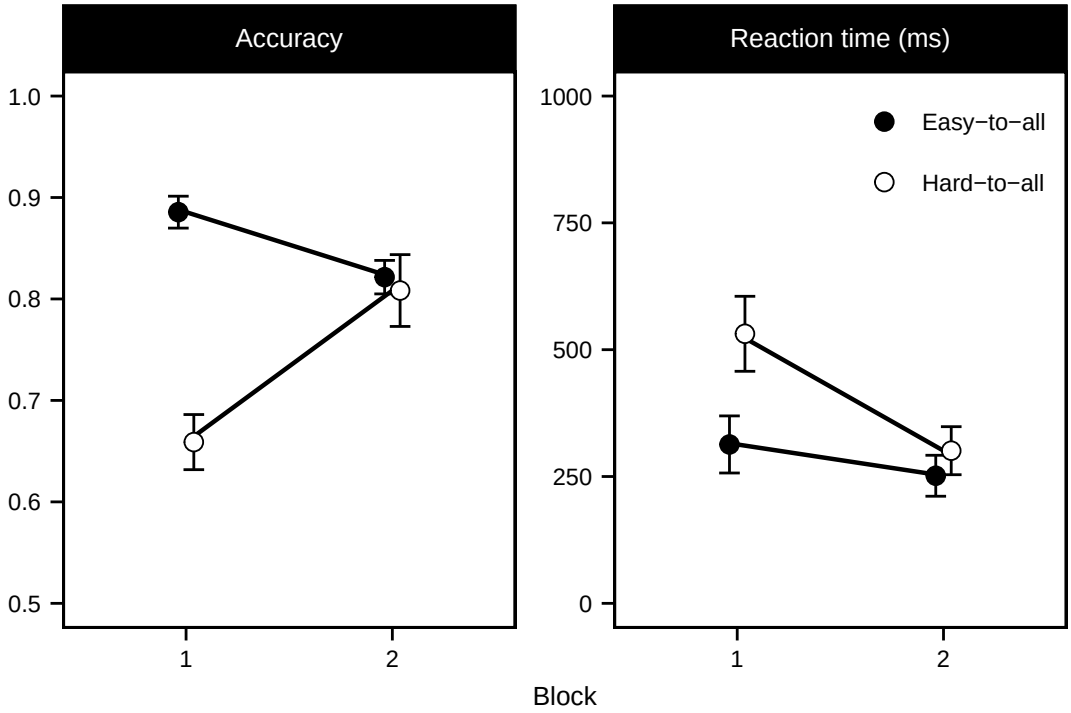


*Figure 3.5:* The average accuracy and reaction time for each block in each condition in Experiment 5. Error bars are difference-adjusted 95% confidence intervals (Baguley, 2012).

the Bayes Factor was below a third, $BF = 0.05$. The Bayes Factor remains in favour of the null hypothesis even if the predicted mean difference is reduced by $3/4$ to $-0.03475$, $SD_{\text{diff}} = 0.0175$. The reaction time data was consistent with this. During Block 1, performance on the hard stimuli was slower than responding on the easy stimul, $t(36) = 3.42$, $d = 1.11$, $p < .001$. However, by Block 2 this difference between conditions had disappeared, $t(36) = 1.16$, $d = 0.38$, $p = .874$.

Additionally, because the participants in each condition saw all the stimuli in Block '2, I was able to examine the difference between the conditions at each level of stimulus difficulty (see Figure 3.6). A mixed ANOVA found a significant main effect of stimulus difficulty, $F(2,72) = 165.83$, $\eta^2 = 0.50$, $p < .001$. As stimulus difficulty increases, average accuracy decreases. Additionally, there was a significant interaction between stimulus difficulty and condition, $F(2,72) = 4.95$, $\eta^2 = 0.03$, $p = .013$. A simple main effects analysis found that the difference between conditions for the easy stimuli approached significance, $t(36) = 1.88$, $p = .068$. However, the difference between conditions did not reach significance for either the moderate, $t(36) = -0.29$, $p = .772$, or the hard stimuli, $t(36) = -0.32$, $p = .754$.

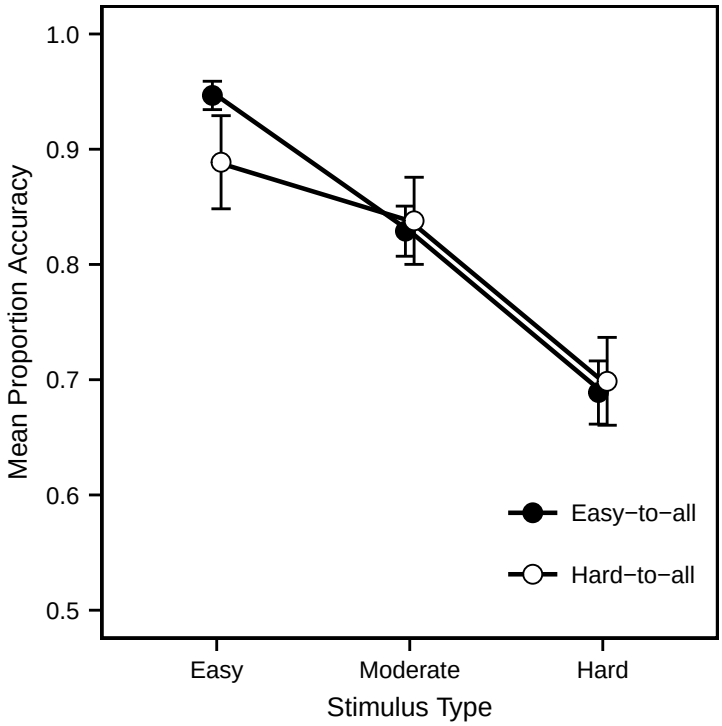


*Figure 3.6:* The average accuracy for each stimulus difficulty level in Experiment 5. Error bars are difference-adjusted 95% confidence intervals (Baguley, 2012).

### 3.4.3 Discussion

Spiering and Ashby (2008) found that participants who were initially trained on a harder version of a classification task had superior performance to participants who were initially trained on the easiest version of the task. However, in Experiments 3 and 4 I failed to find a difference in performance on the moderately difficult stimuli between participants who had been trained on the hardest stimuli and those who had been trained on the easiest stimuli. Experiment 5 aimed to test whether I failed to find a difference in these earlier experiments between conditions because of a ceiling effect. To do this, the participants in both conditions were tested on all the stimuli, as in Block 4 of Spiering and Ashby, rather than just the moderately difficult stimuli, as in Experiments 3 and 4. Additionally, the stimuli were only presented for 350ms. This was to increase the difficulty of the task as a pilot study indicated that just including all the stimuli was not sufficient to reduce performance to levels similar to those found by Spiering and Ashby. These changes were successful in bringing Block 2 performance within the range of performance found in Spiering and Ashby: performance was around 80% for both conditions in the current experiment, whereas for Spiering and Ashby the participants in the hard-to-easy condition scored around 90% and those in the easy-to-hard condition scored around 75%. However, this

experiment still failed to find a difference in final performance between participants initially trained on the easy stimuli and those trained on the hard stimuli.

As all the participants saw all the stimuli, I could look to see whether there was any interaction between stimulus difficulty and condition. Here I found a small effect for the easy stimuli consistent with TAC: performance was slightly higher for participants in the easy-to-all condition than the hard-to-all condition. One argument that this was not an example of TAC comes from the fact that participants in the easy-to-hard condition had seen the easy stimuli before, so this effect may just be an effect of stimulus familiarity. However, speaking against this, there was no such advantage for the hard stimuli in the hard-to-all condition compared to the easy-to-all condition. Therefore, it seems as if any effect of stimulus familiarity may be small and inconsistent (i.e. varying across stimulus types).

## 3.5 Experiment 6

Spiering and Ashby (2008) reported a single experiment that found that hard-to-easy training resulted in better learning of an information-integration category structure than easy-to-hard training. However, Experiment 3 failed to replicate this finding. Indeed, this experiment found evidence for the null hypothesis that there was no difference in performance between the two conditions. Furthermore, I ruled out two possibilities as to why we may have failed to find the original effect. In Experiment 4, I changed the feedback from counterintuitive tones to visual statements. In Experiment 5, participants were tested on all the stimuli, which were only shown from 350ms, to rule out the possibility of a ceiling effect. Nonetheless, both these experiments still found evidence for the null: training order had no effect on final performance. Although Spiering and Ashby (2008) reported an effect of training order on information-integration category learning, three experiments have failed to replicate this effect and instead have even found evidence for the null. These experiments indicate that the original effect may have been an example of a false positive (Type I error).

That being said, there is a possibility that the effect reported by Spiering and Ashby (2008) is a real, but meaningless, effect. Perhaps participants need to complete three blocks of training in order to find a final difference at Block 4. Therefore, in Experiment 3.5 I report the results of a full replication of Experiment 1 of Spiering and Ashby.

Note that, even if Experiment 4 replicates the results found by Spiering and Ashby (2008),

the result would still be uninterpretable due to the confound between primacy and recency. Recall that Spiering and Ashby based their predictions on the type of stimuli the participants first saw in training, i.e. the easy stimuli in the easy-to-hard condition and the hard stimuli in the hard-to-easy condition. However, there is a possibility that participants are basing their responses in the final block on the block they have just seen, i.e. the hard stimuli in the easy-to-hard condition and the easy stimuli in the hard-to-easy condition. This cannot be ruled out in the four block design. However, if there were no differences between conditions at Block 4 it would strengthen the case for Experiment 1 of Spiering and Ashby being a false positive.

### 3.5.1 Method

#### Participants

The participants were 55 undergraduate psychology students recruited from the University of Exeter participation pool. They were randomly assigned to either the easy-to-hard (N=27) or hard-to-easy (N=28) condition. They received 2 research credits in exchange for their participation.

#### Category structure and stimuli

The stimuli were sine-wave gratings displayed on a grey background that were identical to those used in Spiering and Ashby (2008) and Experiment 3. The stimuli used are shown in Figure 3.2.

#### Materials

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

#### Procedure

Participants were tested in individual testing booths and asked to focus on accuracy of responding. The experimental procedure was identical to that in Experiment 3, however two extra training blocks were added. There were 4 blocks of 150 trials each, resulting in a total of 600 trials. In the three training blocks, each of the stimuli were presented in a random order 15 times. The order in which the blocks were presented depended on the condition to which the participants were assigned. In the easy-to-hard condition, participants were shown only the easy stimuli, far from the category boundary in Block 1,

the stimuli of medium difficulty in Block 2 and the hard stimuli, close to the category boundary, in Block 3. In the hard-to-easy condition, the training blocks were shown to participants in the opposite order. Block 4 in both conditions showed all the stimuli in a random order, 5 times each, with feedback. Up to this point, this experiment is identical to that reported by Spiering and Ashby (2008).

After completing the experiment, participants were asked to complete the strategy questionnaire.

**Analysis**

All data analyses were conducted in R (R Core Team, 2015). All trials for which the reaction time was greater than 5000ms were removed.

In addition to the Bayesian analyses conducted on Blocks 1 and 2 in the experiments above, here it is also necessary to look at Block 4. In Block 4, I assumed a two-tailed normal distribution with a predicted mean difference of -0.15, and predicted standard deviation of 0.075. These values were estimated from the results presented in Spiering and Ashby (2008).

### 3.5.2 Results

The average accuracy for each block across the experiment is shown in Figure 3.7. The trial-level raw data will be available at `www.willslab.co.uk/ply75` once this experiment is published. Three participants were excluded from the hard-to-easy condition because they scored below 0.3 for the majority of the experiment[†]. This resulted in 27 participants in the easy-to-hard condition and 25 in the hard-to-easy condition.

For this experiment, the Bayes Factors were calculated using the same technique and prior as described in Experiment 3 for the first two blocks. Here, for Block 1, the sample mean difference was 0.184, with a sample standard error of 0.044. For Block 2, the sample mean difference was 0.01, with a sample standard error of 0.039. The Bayes Factor for Block 4 was calculated using the prior defined in Section 3.5.1. The sample mean difference for Block 4 was -0.049, with a sample standard error of the difference of 0.029.

Following Spiering and Ashby (2008), I compared the mean differences between the conditions at both Block 2 and Block 4. These blocks are where the participants in both

[†]One participant throughout the experiment and two in Blocks 2 to 4

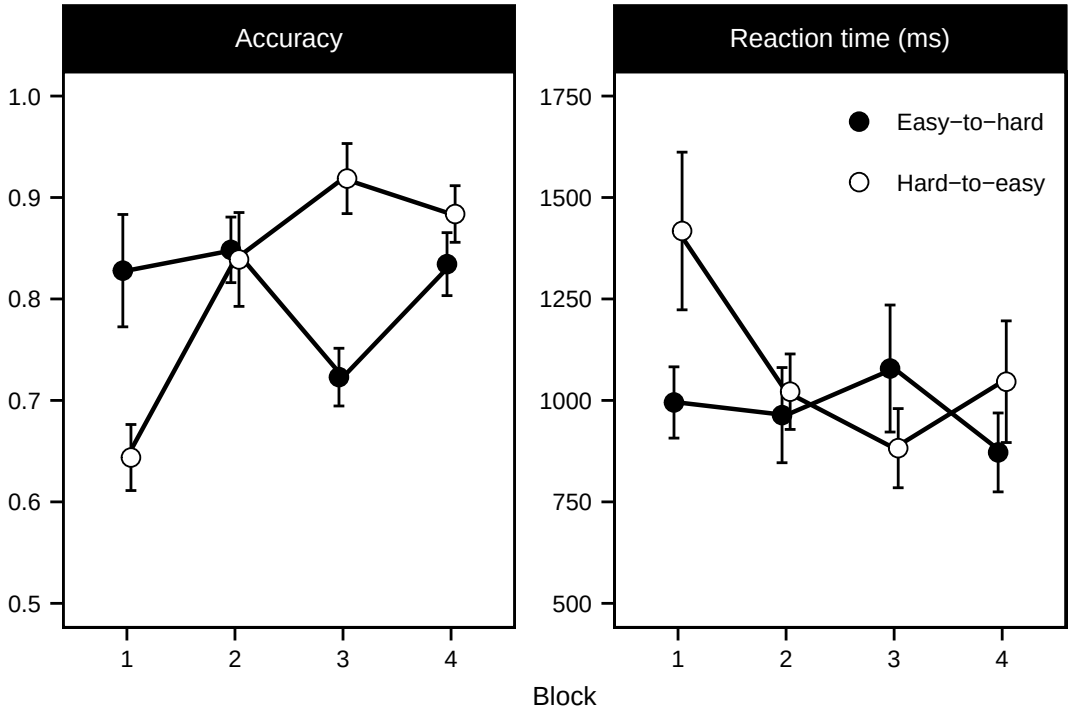


*Figure 3.7:* The average accuracy and reaction time for each block in each condition in Experiment 6. Error bars are difference-adjusted 95% confidence intervals (Baguley, 2012).

conditions saw the same stimuli and so their performance can be compared fairly. The first part of the analysis considers the first two blocks as these are the blocks that can be clearly interpreted without confounding primacy and recency.

During Block 1, performance on the easy stimuli was better than performance on the hard stimuli, $t(50) = 4.09$, $d = 1.13$, $p < .001$, $BF = 636$. However, this initial difference in accuracy had no effect on learning performance in Block 2, $t(50) = 0.25$, $d = 0.07$, $p = .806$. Indeed there was substantial evidence for the null as the Bayes Factor was below a third, $BF = 0.09$. The Bayes Factor remains in favour of the null even if the predicted mean difference is reduced by a half to $-0.070$, $SD_{\text{diff}} = 0.035$.

This pattern of results is also supported by the reaction time data. During Block 1, performance on the hard stimuli was slower than performance on the easy stimuli, $t(50) = 2.96$, $d = 0.82$, $p = .005$. However, by Block 2 this difference had disappeared, $t(50) = 0.56$, $d = 0.15$, $p = .581$.

The data in the first two blocks are consistent with the findings of Experiments 3-5. However, to ascertain whether Spiering and Ashby (2008) was an example of a Type I error, it is also important to see whether there was any difference in performance at Block 4.

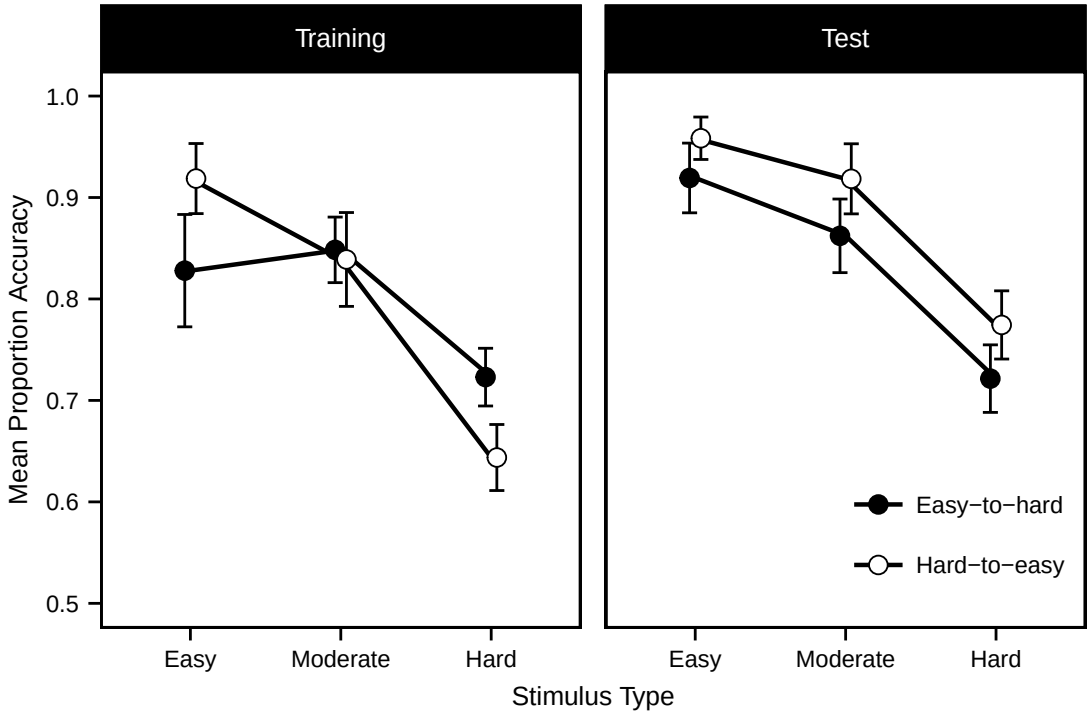


*Figure 3.8:* Average accuracy for each condition for each stimulus type in the Training (Blocks 1 to 3) and Test (Block 4) phases of the experiment. Error bars are 95% difference-adjusted confidence intervals (Baguley, 2012).

In Block 4, performance in the hard-to-easy condition was higher than performance in the easy-to-hard condition, $t(50) = 1.71$, $d = 0.48$, $p = .093$. This is consistent with Experiment 1 of Spiering and Ashby. However, the Bayes Factor was $0.71$. This value indicates a slight preference towards the null hypothesis, but should be interpreted as the test was insensitive.

Again, because all the participants saw all the stimuli, it is possible to compare accuracy for the different stimulus types in each condition as in Spiering and Ashby (2008) and the average performance across Blocks 1 to 3 and Block 4 are shown in Figure 3.8. The performance across training is only shown for completeness as this data is difficult to interpret because different participants have seen different stimuli at each data point. I conducted an ANOVA between condition and stimulus type on the test phase data. Here, the relevant contrasts are Huynh-Feldt corrected as Mauchly's test of sphericity was significant, $W = 0.77$, $p = .002$. The main effect of difficulty was statistically significant, $F(2,100) = 143.69$, $\eta^2 = 0.34$, $p < .001$. As one might expect, the easy stimuli were categorised with higher accuracy than the difficult stimuli. However, the main effect of condition did not reach significance, $F(1,50) = 2.94$, $\eta^2 = 0.05$, $p = .093$ and neither did the interaction, $F(2,100) = 0.30$, $\eta^2 = 0.00$, $p = .705$.

### 3.5.3 Discussion

In Spiering and Ashby (2008), they found that participants learned an information-integration category structure better if they had first been trained on the difficult stimuli compared to those who were first trained on the easy stimuli, but that training order had no impact on learning a conjunction rule-based category structure. Spiering and Ashby argued that this finding supports the dual-system model of category learning COVIS. However, the two experiments reported by Spiering and Ashby both included a confound between primacy and recency. Although the authors argued that the effect in Block 4 was due to the stimuli the participants saw in the initial training block (Block 1), it is not possible to rule out the possibility that the effect was due to the stimuli they saw just prior to test (Block 3).

In Experiments 3, 4 and 5, I included only two blocks and failed to find any evidence that initial training affected final learning performance. Rather I used Bayesian techniques and found evidence for the null hypothesis. These experiments raised the possibility that the effect reported in Spiering and Ashby was an example of a Type I error. However, these three experiments do not rule out the possibility that the original effect was real but meaningless, as it is possible that the confound is required in order to achieve the effect. To address this possibility, in Experiment 4 I replicated all four blocks of Experiment 1 of Spiering and Ashby.

Experiment 6 appears to confirm that the original Block 4 effect, reported in Experiment 1 of Spiering and Ashby (2008), is real but due to the confound between primacy and recency. In the first half of the experiment, which is not confounded, confirms the results I found in Experiments 3, 4 and 5: although there is a significant difference in performance between conditions in Block 1, this effect disappears by Block 2.

In the second half of the experiment, I found a significant difference in performance between conditions in Block 3 in the expected direction. However, in Block 4 there was a difference in the direction predicted by Spiering and Ashby (2008): participants in the easy-to-hard condition performed worse than those in the hard-to-easy condition. Interpreting this difference is difficult because the conditions vary on the stimuli seen in Block 1, the stimuli seen Block 3 and the order of stimuli difficulty over time. However, as there was no difference in performance between the conditions in Block 2, the small Block 4 effect is most likely driven by the stimuli that the participants saw in Block 3. This would be an example of TAC: training on the easy stimuli (for those in the hard-to-easy

condition) in Block 3 results in better performance in Block 4 than training on the hard stimuli (for those in the easy-to-hard condition).

This conclusion raises the question: why did I fail to find TAC in Block 2? One possibility is that participants need to perform at a certain level of accuracy before demonstrating TAC. Performance for both the easy and hard stimuli were lower in Block 1 than they were in Block 3, presumably because the participants had learned something about the task procedure.

Another possibility is that participants need more familiarity with the stimuli in this experiment in order to demonstrate TAC. Other demonstrations of TAC have found the effect using much fewer trials (Church et al., 2013; Suret & McLaren, 2003). However, the stimuli used in these studies only varied systematically across the dimension consistent with the discrimination they were trying to learn. For example, Church et al. varied the speed of birdsong and participants were asked to judge whether the stimulus was fast or slow. All other aspects of the birdsong between the slow and the fast exemplars remained the same. In the current experiments, the stimuli varied systematically in two directions: one consistent with (or parallel to) the discrimination, and one inconsistent with (or perpendicular to) the discrimination. Therefore, it may be more difficult in the current studies to determine the dimension or discrimination and thus, delay participants demonstrating TAC until they understand this.

## 3.6 Strategy analyses

### 3.6.1 Model-based strategy analysis

The evidence above indicates that the Block 2 effect reported in Experiment 1 of Spiering and Ashby (2008) was a false positive and that the Block 4 effect is more consistent with TAC than with the predictions of the dual-system model COVIS. However, there is one remaining objection that a proponent of COVIS might raise. It is possible that category learning is mediated by dual-systems of learning, but that the majority of participants in these experiments did not switch to the optimum system for the category structure. If this was the case, failing to find a dissociation would be predicted by the COVIS model.

This is a valid criticism due to the logic of COVIS experiments (Ashby & Maddox, 2011). COVIS predicts that there are two learning systems that mediate different kinds of learning strategy: rule-based and information-integration (Ashby et al., 1998, 2011). There-

fore, because these systems implement different strategies, the systems will be able to optimally learn different types of category structure. However, the experimental evidence that supports COVIS uses this information the other way around. In the majority of COVIS experiments, the experimenters manipulate the category structure hoping that participants will use the strategies, and thus the system, that is most appropriate to that category structure. However, there is always the possibility that when learning an information-integration category structure participants fail to switch from the Verbal System to the Procedural System.

To overcome this objection, experimental work within the COVIS literature uses a model-based strategy analysis to determine which strategy the participant is using (e.g. Maddox et al., 2003). This strategy analysis is based on General Recognition Theory (GRT; Ashby & Gott, 1988) and is a multi-dimensional generalisation of Signal Detection Theory (Macmillan & Creelman, 2005). For each participant, this analysis determines the optimum decision boundary in stimulus space that separates the stimuli judged by each participant to be in Category A from those in Category B. Each participant is then assigned a strategy type on the basis of characteristics of their optimum boundary. Typically, it is assumed that the category type manipulation has successfully resulted in a change of category learning system if more participants are using the optimum decision bound model for the category structure they have been assigned to than are using a sub-optimum strategy. Although it is important to note that this criterion can be flexible (see Smith et al., 2015, for an example of an alternative criterion).

The GRT-based analysis determines which of a set of experimenter-selected decision-bound models best describes the pattern of responding for each participant (Maddox & Ashby, 1993). Although this analysis is ubiquitous in the experimental COVIS literature, the types of strategy models included, and their precise specifications, often vary between applications of this analysis (see Chapter 5 for more discussion of the effects of this). Therefore, to facilitate comparison with the analysis presented in Spiering and Ashby (2008), here we will use the analysis as reported in that paper.

As in Chapter 2, the set of models considered by Spiering and Ashby (2008) included three main types: rule-based, information-integration and random models. Within the COVIS framework, the unidimensional and conjunction models are considered to represent explicit, rule-based strategies, while the diagonal (GLC) strategy is considered to

represent an implicit, information-integration strategy. Therefore, in these experiments the category structure manipulation will be considered successful if more participants are found to be using the diagonal strategy than a rule-based strategy (either unidimensional or conjunction).

The strategy models used in this analysis were specified as follows:

The *unidimensional* models assume that the participant determines a criterion along one of the stimulus dimensions, either orientation or length. They then make a decision about the category membership of each stimulus by comparing the appropriate stimulus attribute with the criterion value. As an example, for length, this corresponds to a rule of the type: 'Assign to Category A if the stimulus is long, or Category B if short'. The unidimensional models have two parameters: the value of the criterion and the variance of internal (criterial and perceptual) noise.

The *conjunction* model assumes that the participants make two judgements, one for each stimulus dimension, and then combine these to make a judgement about category membership. The conjunction rule in the current analysis was of the type: 'Assign to Category A if the stimulus is short and upright, otherwise assign to Category B'. The conjunction model had three parameters: the two criterion values and internal noise.

The *general linear classifier (GLC)* model assumes that the decision boundary between the categories can be described by a straight line that can vary in gradient and intercept. The unidimensional models are therefore special cases of the GLC model. The GLC model has three parameters: the intercept and slope of the decision bound, plus noise.

There are two random models that assume that participants are responding randomly. The *random* model assumes that participants have no preference for either category: it has no parameters. The *random bias* model assumes that participants respond randomly but prefer one category over the other. It has one parameter that represents the amount of bias.

For each participant, the fit of each of these models was calculated using the Bayesian Information Criterion (BIC; Schwarz, 1978)

$$BIC = r \ln N - 2 \ln L \tag{3.1}$$

where $r$ is the number of parameters in the model, $N$ is the sample size and $L$ is the like-

lihood of the model given the data. The results from this analysis, which was performed using the `grt` package in the R environment (Matsuki, 2014), are reported in Table 3.1.

*Table 3.1:* The proportion of participants that were assigned to each strategy according to the model-based strategy analysis based on the responses from each block for each experiment.

| Condition | Strategies *(wBIC)* | | | | |
|---|---|---|---|---|---|
| | GLC | CJ | UD | RND | BIAS |
| Experiment 3 | | | | | |
| Easy-to-moderate | | | | | |
| Block 1 | 0.41 *(0.76)* | 0.41 *(0.98)* | 0.18 *(0.80)* | - | - |
| Block 2 | 0.82 *(0.96)* | - | 0.18 *(0.74)* | - | - |
| Hard-to-moderate | | | | | |
| Block 1 | 0.60 *(0.95)* | - | 0.20 *(0.80)* | 0.15 *(0.59)* | 0.05 *(0.63)* |
| Block 2 | 0.65 *(0.97)* | 0.10 *(0.75)* | 0.15 *(0.77)* | 0.10 *(0.79)* | - |
| Experiment 4 | | | | | |
| Easy-to-moderate | | | | | |
| Block 1 | 0.35 *(0.80)* | 0.55 *(0.98)* | 0.10 *(0.81)* | - | - |
| Block 2 | 1.00 *(0.99)* | - | - | - | - |
| Hard-to-moderate | | | | | |
| Block 1 | 0.61 *(0.98)* | 0.13 *(0.65)* | 0.17 *(0.64)* | 0.04 *(0.87)* | 0.04 *(0.31)* |
| Block 2 | 1.00 *(0.99)* | - | - | - | - |
| Experiment 5 | | | | | |
| Easy-to-all | | | | | |
| Block 1 | 0.17 *(0.84)* | 0.56 *(0.93)* | 0.28 *(0.75)* | - | - |
| Block 2 | 0.89 *(0.93)* | - | 0.11 *(0.60)* | - | - |
| Hard-to-all | | | | | |
| Block 1 | 0.60 *(0.88)* | - | 0.30 *(0.79)* | 0.10 *(0.73)* | - |
| Block 2 | 0.80 *(0.94)* | - | 0.15 *(0.69)* | 0.05 *(0.81)* | - |
| Experiment 6 | | | | | |
| Easy-to-hard | | | | | |
| Block 1 | 0.44 *(0.91)* | 0.11 *(0.66)* | 0.41 *(0.74)* | 0.04 *(0.89)* | - |
| Block 2 | 0.74 *(0.98)* | 0.04 *(0.41)* | 0.19 *(0.73)* | 0.04 *(0.89)* | - |
| Block 3 | 0.63 *(0.97)* | - | 0.30 *(0.80)* | 0.04 *(0.90)* | 0.04 *(0.74)* |
| Block 4 | 0.78 *(0.96)* | - | 0.19 *(0.84)* | 0.04 *(0.87)* | - |
| Hard-to-easy | | | | | |
| Block 1 | 0.43 *(0.88)* | - | 0.29 *(0.74)* | 0.21 *(0.69)* | 0.07 *(0.51)* |
| Block 2 | 0.64 *(0.95)* | 0.18 *(0.73)* | 0.11 *(0.71)* | 0.07 *(0.73)* | - |
| Block 3 | 0.50 *(0.92)* | 0.11 *(0.62)* | 0.36 *(0.56)* | 0.04 *(0.60)* | - |
| Block 4 | 0.86 *(0.97)* | 0.04 *(0.50)* | 0.07 *(0.78)* | 0.04 *(0.76)* | - |

Strategies: GLC=General linear classifier, CJ=Conjunction, UD=Unidimensional, RND=Random.

Although not typically a part of the standard model-based strategy analysis used in the COVIS literature (although see Roeder & Ashby, 2016, for a Bayes Factor approach), it is also informative to look at the performance of the best-fitting model relative to the

performance of the competing models. If the winning model performs much better than its competitors, i.e. it fits better to the data, we can be more confident that this model provides the best description of the participant's behaviour from among the pre-specified alternatives. On the other hand, if the winning model performs only slightly better than the alternatives, our confidence that the winning model best describes the participant's responses should be lower. There are several cases where this might occur. For example, the participant may be swapping between strategies, applying a single strategy inconsistently with lapses in attention or even using a strategy not included within the set of models the analysis can select from (Donkin et al., 2015). Therefore, it is important to investigate the fit of the strategy models.

One principled way of evaluating the validity of the model-based analysis is by calculating Schwarz weights (Wagenmakers & Farrell, 2004). Schwarz weights ($w_i(BIC)$) are defined as the probability that model $i$ is best, in terms of minimising the BIC, given the data and the set of competing models. The average Schwarz weights for the winning models are included in 3.1. From these, it is also possible to calculate the normalised probability that the optimum diagonal strategy is preferred over rule-based strategies (i.e. conjunction and unidimensional) for each participant. From the Schwarz weights, the normalised probability that the diagonal strategy model is to be preferred over the conjunction and unidimensional strategy models is calculated using:

$$\frac{w_{GLC}(BIC)}{w_{GLC}(BIC) + w_{CJ}(BIC) + w_{UD}(BIC)} \tag{3.2}$$

where $w_{GLC}(BIC)$, $w_{CJ}(BIC)$ and $w_{UD}(BIC)$ are the Schwarz weights for the diagonal, conjunction and unidimensional strategy models respectively. These values are reported for each experiment in Table 3.2.

Table 3.1 shows that a majority of participants in the test blocks (Block 2 and 4) of each condition in each experiment were found to be using the optimum diagonal strategy for the category structure. Furthermore, Table 3.2 shows that the probability of the 'best' model being the diagonal general linear classifier by the second block is high in every experiment. According to the logic of this analysis as used in the COVIS literature, this indicates that participants are capable of using the Procedural System to learn this category structure. Therefore, the failure to find evidence of initial training on final performance cannot be attributed to participants never being able to switch to the optimum learning system

*Table 3.2:* The normalised probability that the optimum diagonal strategy is preferred over rule-based strategies.

| Condition | Experiment | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Easy first | | | | |
| Block 1 | 0.29 | 0.34 | 0.21 | 0.49 |
| Block 2 | 0.99 | 0.82 | 0.85 | 0.77 |
| Block 3 | x | x | x | 0.66 |
| Block 4 | x | x | x | 0.77 |
| Hard first | | | | |
| Block 1 | 0.71 | 0.64 | 0.59 | 0.44 |
| Block 2 | 0.99 | 0.69 | 0.78 | 0.67 |
| Block 3 | x | x | x | 0.54 |
| Block 4 | x | x | x | 0.85 |

for the task.

**A warning about interpreting the Block 1 strategies**

Looking carefully at the data from Tables 3.1 and 3.2 it may also be tempting to draw conclusions about the interaction between Training Order Condition, Block and the proportion of participants using the optimum strategy. For instance, looking at the entries for Block 1 in Table 3.1, you can see that more participants who were initially shown the hard stimuli were found to be using the optimum strategy than participants who were initially shown the easy stimuli. Arguably, this would support the predictions of COVIS: that participants in the hard first conditions switch faster to the Procedural System and can thus implement the optimum diagonal strategy faster. Unfortunately, the model-based analysis cannot be used in this way.

This comparison is problematic because the results of the strategy analysis are conditional on the category structure it is applied to. Here, the issue lies in the fact that for the easy stimuli the unidimensional strategy model would result in similar performance to the optimum diagonal strategy model. Whereas for the hard stimuli, the unidimensional strategy would score much lower than the optimum diagonal strategy. This issue is exaggerated as the unidimensional strategy model has two parameters whereas the diagonal general linear classifier strategy model has three. Therefore, if the fit is similar between these two models the BIC will always favour the simplest. This biases the analysis to-

wards finding more unidimensional strategies in Block 1 of the initially easy conditions than the initially hard conditions. This issue will be considered in much greater depth in Chapter 5.

### 3.6.2 Verbal report analysis

The model-based strategy analysis is at best a secondary measure of whether a participant's strategy was implicit; the diagonal (GLC) strategy is assumed to be a marker of implicit responding, without ever asking participants. Therefore, one should be careful in assuming that the participants who were identified as using the optimum diagonal strategies were learning the category structure implicitly. To check whether or not this was the case, I asked participants to describe the strategies they used in three of the four experiments reported above (participants in Experiment 4 were not given the verbal report questionnaire due to experimental error). These questionnaires were independently coded by the author (CERE) and a student volunteer (GW). First, each verbal report was examined to determine whether the participant had reported an explicit categorisation strategy or not. Second, the available strategy descriptions were sorted into groups of four main kinds: information-integration, two-dimensional, conjunction and unidimensional. The inter-rater reliabilities for these initial codings are reported in Table 3.3. Then, any discrepancies were easily resolved through discussion with reference to the strategy descriptions below.

Across the three experiments, five types of strategy were identified:

Participants were placed in the *implicit* group if they described attempting to make the stimulus dimensions commensurable, such as 'Stimuli for which the line was longer than it was upright should be assigned to category A' or if they said anything that could be reasonably interpreted as a statement that they based their classification on overall similarity. Note that overall similarity descriptions are commonly found in other studies, not within the COVIS-framework, in which our lab has elicited verbal reports (e.g., Wills et al., 2013). Although, note that such studies do not support the conclusion that overall similarity classification is implicit.

Participants were placed in the *complex rule* group if they described a rule using both stimulus dimensions in a complicated fashion. Example strategies include rule-plus-exception strategies such as "upright stimuli were in Category A and flat stimuli in Category B. However, if the stimulus was upright and had very few bars it was in Category B"

*Table 3.3:* Summary of the inter-rater reliability statistics for judging whether the participant had a strategy and the type of strategy identified. Also, listed are the number of participants in each condition in each experiment that failed to report a strategy.

| Experiment | Explicit strategy | | Strategy Type | | Removed | |
|---|---|---|---|---|---|---|
| | $\kappa$ | $p$-value | $\kappa$ | $p$-value | Easy first | Hard first |
| Experiment 3 | 0.88 | <.001 | 0.52 | .001 | 1 | 2 |
| Experiment 5 | 0.66 | <.001 | 0.72 | .001 | 1 | - |
| Experiment 6 | 1.00 | <.001 | 0.90 | .001 | 2 | 3 |

or sequential unidimensional rules such as “upright stimuli were in Category A and flat stimuli were in Category B. For stimuli at 45 degrees, it was in Category A if it had less than three bars and Category B if it had more than three bars.”

Participants were placed in the *conjunction* group if they used both stimulus dimensions and described categorising stimuli using a logical conjunction rule such as ‘upright stimuli with lots of lines were in Category A, otherwise they were in Category B.’

Participants were placed in the *two-dimensional* group if they described using both stimulus dimensions but with descriptions that were too unclear to be assigned to more specific categories.

Participants were placed in the *unidimensional* group if they described categorising stimuli based solely on either bar frequency (or line length in Experiment 4) or stimulus orientation.

Several participants were removed from each experiment because they described abstract elements of the experimental setup, such as which buttons to press, rather than their sorting strategy and so were excluded. The number of excluded participants in each condition in each experiment are also displayed in Table 3.3.

As can be seen in Table 3.4 the majority of participants reported using explicit complex rules to solve the categorisation task. This is inconsistent with the predictions of COVIS. This model would predict that a non-trivial number of participants would report using implicit strategies. Furthermore, that more participants would report using implicit strategies in the hard-to-easy condition than in the easy-to-hard condition. This is quite clearly not the case.

*Table 3.4:* The proportion of participants that reported using each strategy type.

| Condition | Verbal reports | | | | |
|---|---|---|---|---|---|
| | Implicit | 2D | CJ | Complex | UD |
| Experiment 3 | | | | | |
| Easy-to-medium | - | 0.11 | 0.17 | 0.67 | 0.06 |
| Hard-to-medium | - | 0.11 | 0.22 | 0.56 | 0.11 |
| Experiment 5 | | | | | |
| Easy-to-all | - | 0.11 | 0.21 | 0.42 | 0.26 |
| Hard-to-all | - | 0.05 | 0.45 | 0.50 | - |
| Experiment 6 | | | | | |
| Easy-to-hard | - | 0.04 | 0.12 | 0.48 | 0.36 |
| Hard-to-easy | - | - | 0.24 | 0.60 | 0.16 |

Strategies: 2D=Complex rule using both dimensions, CJ=Conjunction, UD=Unidimensional.

Additionally, Table 3.4 shows evidence consistent with a single-system rule-based approach. In two of the experiments, at the end of the experiment more participants report using a unidimensional strategy in the easy-to-hard conditions than in the hard-to-easy conditions.

It is also interesting to compare the strategies reported by participants with those identified by the model-based strategy analysis. In Table 3.5 the strategies each participant reported using are compared with the ones they were assigned using the strategy analysis. This table shows that all participants who were found to be using a diagonal strategy by the model-based analysis reported using a strategy reported using a rule-based strategy. This is contrary to the predictions of both COVIS and the model-based analyses. These would predict that participants would report using information-integration strategies or use other explanations that sounded more implicit in nature.

## 3.7 General discussion

This chapter looked at how initial training difficulty impacts category learning. Spiering and Ashby (2008) argued that initial training on a hard discrimination resulted in better performance than initial training on an easy discrimination. They explained this finding in terms of the COVIS model of category learning. Spiering and Ashby argued that participants in the easy-to-hard condition could perform very well on these easy stimuli by using a sub-optimum unidimensional strategy and so would delay swapping from the explicit, Verbal System to the optimum Procedural System. In contrast, participants who were first

*Table 3.5:* Comparison of the strategies assigned to each participant in the model-based strategy analysis with those they reported.

| Model-based | Verbal reports | | | | |
|---|---|---|---|---|---|
| | Implicit | Complex | CJ | 2D | UD |
| Experiment 3 | | | | | |
| GLC | - | 16 | 7 | 1 | 1 |
| CJ | - | 1 | - | 1 | - |
| UD | - | 3 | - | 1 | 2 |
| RND | - | 1 | - | - | - |
| Experiment 5 | | | | | |
| GLC | - | 17 | 10 | 2 | 2 |
| CJ | - | - | - | - | - |
| UD | - | 1 | 2 | - | 2 |
| RND | - | - | - | 1 | - |
| Experiment 6 | | | | | |
| GLC | - | 25 | 3 | 1 | 6 |
| CJ | - | - | 4 | - | 2 |
| UD | - | 2 | 2 | - | 4 |
| RND | - | - | - | - | 1 |

Strategies: 2D=Complex rule using both dimensions, CJ=Conjunction, UD=Unidimensional.

shown the difficult stimuli would swap to the optimum implicit system much sooner as it would be clear that rule-based approaches were not working.

These results could also be explained by two simpler theoretical accounts. The first account posits a single, rule-based system of category learning. This approach predicts that final learning performance is different between conditions because of the type of rule initially learned by each participant. Those who were initially trained on the easy stimuli are likely to learn a simple rule that generalises poorly to the more difficult stimuli. Those who were initially trained on the hard stimuli are likely to learn a more complex rule that generalises to the other stimuli well. This means that participants who were initially trained on the easy stimuli perform worse at final test than those trained on the more difficult stimuli.

A second possibility cannot be ruled out due to a confound in Spiering and Ashby's (2008) experimental design. Spiering and Ashby argued that the effect of training order on information-integration category learning is due to the type of stimuli the participant saw *initially*. However, the stimuli in the training block just prior to test are hard in the easy-to-hard condition and easy in the hard-to-easy condition. Therefore, it is possible that participants' performance in the test phase following training reflects the most *recent* rather than the initial training. If this were the case, their results would actually be similar to TAC. In TAC, initial training on an easy version of a discrimination (compared to training on a hard version) results in superior performance on the hard discrimination at test (Lawrence, 1952). Therefore, their Block 4 effect could alternatively be ascribed to TAC rather than the interaction of two systems of category learning.

This chapter reported four experiments, all of which failed to replicate the effect reported by Spiering and Ashby (2008). In Experiment 3, I re-examined only the first two blocks of Spiering and Ashby to avoid the possibility of confounding the effect of initial training with the effect of most recent training. In contrast to Spiering and Ashby, I found evidence for the null hypothesis: the type of initial training had no effect on final performance. In Experiment 4, I again compared easy-to-moderate with hard-to-moderate training. Although, for this experiment I changed the feedback and stimuli to rule out the possibility that they were confusing the participants. Here, I again found evidence for the null hypothesis. In Experiment 5, I examined the possibility that Experiments 3 and 4 were subject to a ceiling effect. To do this, Block 2 included all the stimuli and the participants

were trained under time pressure. Once again, I failed to find a differential effect of initial training. Finally, in Experiment 6, I replicated the entirety of Spiering and Ashby (2008). Here, in Block 2, I once again found evidence for the null hypothesis. However, in Block 4 participants in the hard-to-easy training condition performed better than the participants in the easy-to-hard training condition. Arguably, this was more consistent with TAC than COVIS as there was evidence for the null in Block 2, so both conditions were starting from the same point in Block 3. Therefore, it is most likely that Block 3 was driving this effect not Block 1.

A possible explanation as to why we failed to find the effect reported in Spiering and Ashby (2008) is that they may have failed to exclude mis-learners (their paper is unclear on this point). Mis-learners are those participants whose performance was markedly different from chance, but in the wrong direction, i.e. they scored around 10% rather than 90%. These participants have obviously learned the features of the category structure, but mistook the response key for each category structure. This is to be compared with non-learners who score around chance (here 50%), who have not learned the features of the category structure. This is a problem for Spiering and Ashby because of the counter-intuitive choice of feedback: the high tone indicated an incorrect response, whereas the low tone indicated a correct response. When I used this feedback (Experiments 3 and 6) several participants mis-learned the category structure.

The mis-learning participants are important because the conclusions one draws from the experiment critically depend on whether or not they are included in the analyses. In Experiment 3, three participants were excluded from the easy-to-moderate condition. However, when they are included the pattern of results is similar to Spiering and Ashby (2008, although it does not reach significance): hard initial training results in better Block 2 performance than easy initially training. In Experiment 6, three participants were removed from the hard-to-easy condition. In this case, any difference between conditions in Block 4 disappears. Therefore, the impact of including mis-learners depends on which condition they were in. Additionally, this raises possibility that the original effect was due to several participants in the easy-to-hard condition mis-learning the category structure and not being excluded.

Not only do these experiments indicate that the conclusions we can draw from them are sensitive to mis-learners, but it also highlights the importance of applying a learning

criterion to experiments within the COVIS literature. This observation is not novel. For example, Zeithamova and Maddox (2006) looked at the effect of concurrent load on learning of rule-based and information-integration category structures. As predicted by the COVIS model, Zeithamova and Maddox found that concurrent load negatively impacted rule-based but not information-integration category learning. However, when Newell et al. (2010) re-examined these experiments, they found that the conclusions they could draw were dependent on excluding non-learners. When non-learners were included in these experiments, the data were consistent with the COVIS model. However, when they were excluded the experiment failed to find evidence of a dual-system model.

The experiments in the current chapter also indicate that a single-system, rule-based model might be sufficient to explain these experiments and the effect reported by Spiering and Ashby (2008). In three of the experiments above, I asked the participants to describe the strategy that they used to complete the category learning task. These descriptions demonstrated that all the participants who reported a relevant strategy were using a rule-based strategy. Additionally, those who failed to report a strategy only did so because they reported the specifics of the experiments (such as which button to press). Of these participants none of them reported anything that could possibly be interpreted as an implicit strategy. Furthermore, the verbal reports provided preliminary evidence that the participants in the initially easy conditions were more likely to report using a unidimensional strategy than those in the initially hard conditions. Together these verbal reports indicate that the participants who are first shown the easy stimuli may be more likely to find a simple rule and stick to it, whereas those first shown the difficult stimuli may find a more complex rule. Contrary to our predictions, these differences in initial strategy selection do not appear to affect performance much by Block 2. However, it is possible that these strategies interact with the number of blocks and the relative success of these strategies in order to produce the predicted effect in Block 4.

However, these conclusions are limited because we do not know how participants' strategies change during training. Although in theory the model-based analysis can be used to determine the pattern of responses for each block, the analysis does not seem consistent with what the participants report doing (also see Chapter 5). In contrast, the verbal reports appear to more accurately identify which type of strategy each participant used. However, they do not provide information about how each participant's strategy changes over time as the report was given after the end of the experiment. Therefore, to deter-

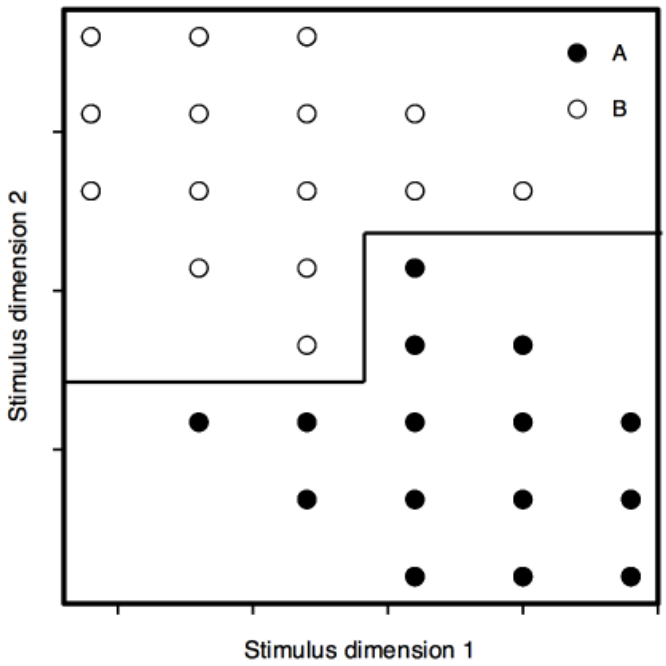


*Figure 3.9:* Suggested category structure to look at the effects of sub-optimum strategy performance in a rule-based category structure.

mine whether the strategies that participants initially use have a large impact on learning, future work might ask participants which strategy they used after each block of the experiment. An alternative approach might interleave no feedback test phases between each block, such as in the experiments reported in Chapter 2 and Ashby et al. (2002). If these test blocks consisted of transfer stimuli that covered the whole of stimulus space, it may improve the accuracy of the model-based analysis (Donkin et al., 2015). Additionally, this would facilitate comparisons between training order conditions as participants in both conditions would be seeing the same stimuli.

Another possibility to be explored is whether differences in participants' strategies could be found in any category structure or whether they are specific to learning the information-integration category structure. If it is the case that it is the relevant performance of sub-optimum strategies that drive the effect in Experiment 1 of Spiering and Ashby (2008), then the conjunction structure may have been a poor choice of control. This is because simpler strategies resulted in the same level of performance as more complicated ones for every level of stimulus difficulty. A better alternative might be the structure shown in Figure 3.9. Here, a unidimensional strategy would score perfectly on the easy stimuli (those furthest from the category boundary) but poorly on the hard stimuli, the same as in the information-integration category structure. This would be something interesting to investigate in future work.

### 3.7.1 Conclusion

The current work challenges Spiering and Ashby's claim that it is sometimes best to start training with the most difficult items. All-in-all, the current work illustrates the dangers of making striking, novel, claims on the basis of any single experiment.

# Chapter 4

# Recognition

In Chapters 2 and 3, I re-examined two results that claimed to support the dual-system model of category learning COVIS. In Chapter 2, I demonstrated that Ashby et al.'s (2002) differential effect of training type on learning was due to a confound between the two category structures they compared. In Chapter 3, I reported experiments that failed to replicate the effect found in Experiment 1 of Spiering and Ashby (2008), indicating that this result also does not support the COVIS model.

Interestingly, the implications of these chapters go beyond merely removing two sources of support for the COVIS model. COVIS assumes that optimal learning of information-integration category structures is mediated by the Procedural System (Ashby & Valentin, 2016). Critically, this system is hypothesised to produce "category knowledge opaque to declarative consciousness" (p. 2476, Smith et al., 2015). Therefore, participants who learn information-integration category structures well should not be able to report how they achieved this. In other words, COVIS predicts that the majority of participants who took part in the experiments reported in Chapters 2 and 3 (i.e. those who learned an information-integration category structure) should have no explicitly-reportable category knowledge.

Whether there was support for this prediction in the previous experiments depends on how "category knowledge" is formalised. Within the COVIS literature, category knowledge is always quantified using the GRT-informed model-based strategy analysis (Maddox & Ashby, 1993). In this analysis, participants who are identified as using a diagonal (GLC) strategy are assumed to have no category knowledge. As the majority of participants in the information-integration conditions above were identified as using a diagonal (GLC) strategy, proponents of COVIS would conclude that these participants were learning the category structure implicitly and thus, had no category knowledge. However, when these same participants were asked to report their category knowledge, a different picture emerged. Nearly all the participants–regardless of the category structure they had

learned and the strategy they were assigned using the GRT analysis–were able to explicitly report a classification strategy.

This contradiction between the result of the GRT analysis and the verbal reports raises the possibility that two of the fundamental assumptions in the COVIS literature may be wrong. First, the results of the model-based analysis may be a poor indicator of the strategies that participants use. This possibility is explored in detail using simulations in Chapter 5. Second, participants may learn information-integration category structures explicitly, using rule-based strategies rather than implicitly, using a procedural learning mechanism. In the current chapter, I investigate this possibility in four experiments that examine participants' recognition memory performance after learning rule-based or information-integration category structures.

## 4.1 Introduction

Despite the "implicitness" of the Procedural System being a fundamental assumption of the COVIS model, there have been no studies that directly investigated whether participants can explicitly access information-integration category knowledge (Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2016). Instead, proponents of COVIS have looked at whether manipulations proposed to affect other procedural learning tasks also affect information-integration category learning. For example, some have examined the effect of switching the response buttons on rule-based and information-integration category learning (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004). They found that participants learning a rule-based category structure were not affected by the manipulation, whereas those learning an information-integration category structure were. As other studies have demonstrated that this is indicative of procedural learning (e.g. Willingham, Wells, Farrell, & Stemwedel, 2000), they generalise from these skill learning paradigms to information-integration category learning (although see Nosofsky et al., 2005, who demonstrated the dissociation was due to the increased complexity of the information-integration structure). Then, as procedural learning is defined as requiring no conscious awareness (Ashby & Crossley, 2012; Squire, 2004), they infer that this must also be true for information-integration category learning.

Another example of research that used this logic investigated feedback timing. Maddox and colleagues (2003; 2005) found that rule-based category learning is unaffected by feedback timing but that information-integration category learning is impaired. As this

has been demonstrated in other areas of the procedural learning literature (Willingham, 1998), they assume this demonstrates procedural learning and thus, *a priori* must be implicit (although see Dunn et al., 2012, who demonstrated the dissociation was due to the choice of mask used).

However, some neuroscientific and behavioural studies suggest that participants may learn information-integration category structures explicitly. The neuroscientific evidence comes from a study by Carpenter et al. (2016) that found greater activation of the medial temporal lobe in information-integration category learning than rule-based category learning. The medial temporal lobe has long been considered critical for explicit memory (e.g., Conroy, Hopkins, & Squire, 2005; Squire, 1992). Therefore, Carpenter et al.'s results suggest that information-integration category learning involves explicit memory processes to a greater extent than rule-based category learning.

That being said, Carpenter et al.'s results are at odds with a previous neuroimaging study of rule-based and information-integration category learning (Nomura et al., 2007). However, Carpenter et al. argue that the differences between their study and that of Nomura et al. are due to methodological problems with the Nomura et al. study. This argument is supported by the fact that Carpenter et al.'s results are broadly in line with the only other closely-related neuroimaging study (Milton & Pothos, 2011). Furthermore, it is also consistent with the following behavioural evidence.

The behavioural evidence that participants learn information-integration category structures explicitly comes from the previous two chapters. In these experiments, the majority of participants who learned information-integration category structures were found by the model-based strategy analysis as using a diagonal (GLC) strategy. Within the COVIS literature, this is taken as evidence that participants are using the implicit Procedural System to learn. However, all of these participants were also able to explicitly report the strategies they used. Furthermore, these strategies were often not optimal. Indeed, no participants reported using a strategy that indicated that they attempted to make the two stimulus dimensions commensurable, as would be supposed for an information-integration category structure (Ashby et al., 1998; Ashby & Valentin, 2016). Rather, they reported using sub-optimum rule-based strategies, such as conjunction or unidimensional rules. Similar results were found in work investigating the weather prediction task, another categorisation task sometimes assumed to be implicit (Knowlton, Mangels, & Squire,

1996), participants were also able to explicitly report the strategies they used (Lagnado, Newell, Kahan, & Shanks, 2006; Newell et al., 2007). Additionally, participants in tasks commonly assumed to be learned implicitly, such as the serial reaction time task, are unable to report even partial task knowledge (Yeates, Jones, Wills, Aitken, & McLaren, 2012, 2013). Therefore, it seems critical to directly test whether participants have explicit access to category knowledge when learning information-integration category structures.

### 4.1.1 Recognition memory

To determine whether participants actually learn information-integration category structures using an implicit procedural learning mechanism as advocated by the COVIS model, the following experiments will examine recognition memory performance after learning rule-based and information-integration category structures. This is for several interconnected reasons. First, recognition memory performance is commonly supposed to be a test of explicit memory processes (Berry, Shanks, Speekenbrink, & Henson, 2012; Gabrieli & Fleischman, 1995). Furthermore, previous evidence has pointed to a strong connection between the medial temporal lobe and recognition memory (e.g., Squire, Wixted, & Clark, 2007). Therefore, a stronger reliance on recognition memory processes in information-integration compared to rule-based category learning would explain the differential activation in the temporal lobe found by (Carpenter et al., 2016).

Second, previous research has found a close relationship between recognition memory performance and exemplar models of category learning (e.g., Medin & Schaffer, 1978; Nosofsky, 1988). Exemplar models bear many similarities to the Procedural System of COVIS. Much like the Procedural System, exemplar models assume that a stimulus is assigned to category for which it is most similar to the exemplars of that category. However, unlike the Procedural System, exemplars are not assumed to be implicitly learned. Indeed, these models have been used to explain various recognition memory effects (e.g., Davis, Love, & Preston, 2012; Nosofsky, 1988; Nosofsky & Zaki, 1998; Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). Furthermore, Hoffmann, von Helversen, and Rieskamp (2014) found that episodic memory (as defined by recall and recognition tasks) was related to performance on a multiplicative category learning task, similar to an information-integration category structure task. Participants who had better episodic memory also performed better on the category learning task*.

*It is interesting to note that Hoffmann et al. (2014) also looked for a relationship between implicit perceptual processes and the categorisation tasks. However, this was not possible as the tasks that were

### 4.1.2 Predictions

I would predict that recognition memory performance would be higher after learning an information-integration category structure than a rule-based structure. This is because in the previous chapters, I found that participants reported using complex rule-based strategies to learn an information-integration category structure. The more complex a rule-based strategy, the more processing is required of each stimulus as it involves comparing the stimulus with more decision boundaries. For example, characterising a conjunction strategy in terms of decision boundaries requires two boundaries whereas a sequential unidimensional rule strategy[†] requires three (for examples of behavioural work demonstrating participants' use of even more complex strategies see Donkin et al., 2015). Using these more complex strategies is thus likely to encourage participants to pay closer attention to the features of the stimulus. Previous evidence has shown that deeper processing of stimuli increases recognition memory performance (e.g., Craik & Tulving, 1975; Gardiner, Java, & Richardson-Klavehn, 1996; Jacoby, Shimizu, Daniels, & Rhodes, 2005). Therefore, I would predict that participants learning information-integration category structures using complex rules would have superior recognition memory performance than participants learning a simple rule-based structure using simple rules.

In contrast, COVIS predicts that information-integration category structures are learned procedurally by non-declarative memory systems (Ashby & O'Brien, 2005). Therefore, COVIS would predict that participants will have very poor recognition memory after learning an information-integration category structure. What is less clear is whether COVIS would predict greater than chance recognition memory for items in a rule-based category structure. COVIS assumes that rule-based category structures are learned by the Verbal System. This system is hypothesised to rely on declarative memory processes and attention and result in explicit knowledge of the category (Nomura et al., 2007). This might suggest that recognition memory performance would be above chance. However, other theories of memory (such as Squire, 1992) might suggest that the systems of working memory and recognition memory, although both declarative memory processes, are distinct and non-overlapping. Therefore, one might not expect recognition memory to be above chance.

hypothesised to tap into implicit perceptual processes were found not to be correlated.

[†]For line stimuli, which vary on length and orientation, this strategy corresponds to the following rule: "If the stimulus is short it is in Category A, if the stimulus is long it is in Category B. If the stimulus is medium-long, then if it is close to upright it is in Category A otherwise it is in Category B."

## 4.2 Experiment 7

In this experiment, recognition memory will be compared after participants have learned either unidimensional or information-integration category structures. In Chapter 2, I argued that a unidimensional rule-based structure is typically a poor control for an information-integration category structure due to the different number of relevant dimensions to make a correct categorisation judgement. Here, however, this flaw may actually be a strength.

The predictions of the COVIS model are built on the assumption that one category structure is easy to verbalise and the other is not. Therefore, for COVIS, either a unidimensional or conjunction category structure should be adequate as they should both be learned by the explicit Verbal System. On the other hand, my predictions are based on the assumption that differences in memory may arise because of the differences in strategies that participants use. A unidimensional category structure should encourage participants to use unidimensional rules, whereas an information-integration category structure should encourage participants to use complex, two-dimensional rules. However, a conjunction category structure would also encourage participants to use a two-dimensional rule. Therefore, using a unidimensional category structure rather than a conjunction would maximise the differences in strategies between the rule-based and information-integration category structures, thus increasing the likelihood that I would find a difference in recognition performance.

Furthermore, this experiment is not looking for an interaction between an experimental manipulation and learning of two different category structures. The lack of a second independent variable means that there will be no issues with that manipulation having a smaller or greater effect depending on the number of dimensions of the category structure (Dunn & Kirsner, 2003). Therefore, the issues raised in Chapter 2 should have little impact.

### 4.2.1 Method

#### Participants

The participants were 42 undergraduate psychology students recruited from the Plymouth University participation pool. They received 2 credits in exchange for their participation.

### Stimuli and category structures

The stimuli were 36 grey squares that varied in brightness and size displayed on a white background. The stimuli seen by each participant depended on which category structure they learned.

Half the participants were randomly assigned to learn a unidimensional rule-based category structure and the other half to learn an information-integration category structure as illustrated in Figure 4.1. The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness (11 participants) or size of the square stimuli (10 participants)–and two information-integration category structures–where the optimum boundary had either a positive (10 participants) or negative gradient (11 participants). The abstract representation of the information-integration positive category structure is identical to that used by Spiering and Ashby (2008) with a row of 6 stimuli added perpendicular to the category boundary to bring the total number of stimuli up to 36. These stimuli were added to facilitate the random selection of a third of stimuli as "unseen" items for the recognition task. The remaining category structures are rotations ($\pi/4$, $\pi/2$, $3\pi/4$ radians) of this original structure around the origin and then translated so that 'centre of gravity' (i.e. the mean of the brightness and size dimensions) of the points remained the same. The abstract stimuli coordinates were log-scaled so that all adjacent stimuli were approximately equally perceptually discriminable. The brightness value corresponded to the proportion brightness of the squares. The size value for each stimulus corresponded to the length of each side of the square in centimetres. The abstract and logged scaled stimulus coordinates are archived along with the raw data at `www.willslab.co.uk/ply40`.

### Materials

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

### Procedure

The experiment was split into three phases: category training, recognition test and finally, category test.

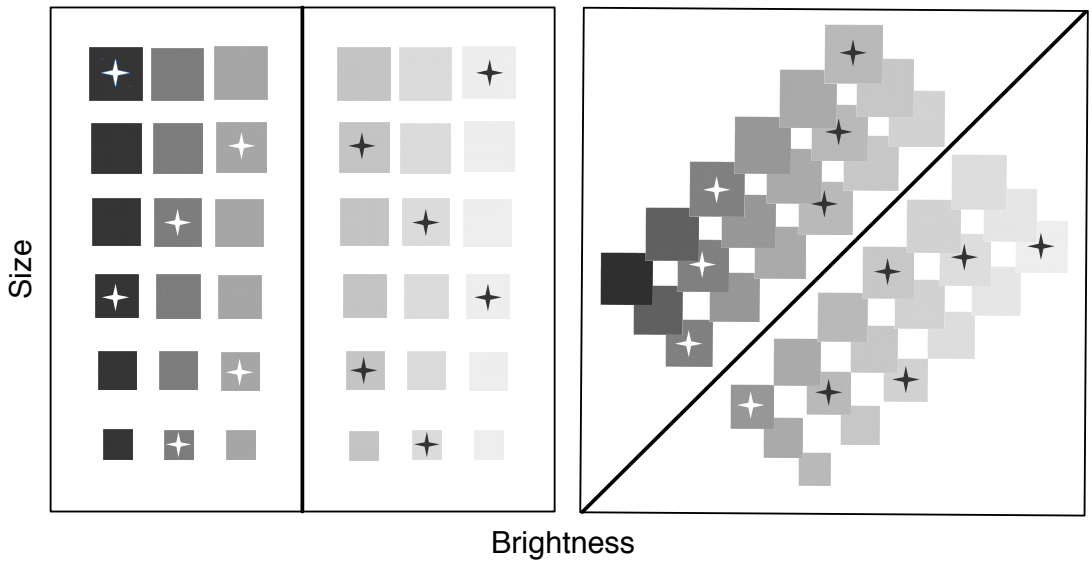


*Figure 4.1:* Examples of stimuli used in one of the unidimensional and one of the information-integration category structures in Experiment 7. The stars mark a possible pattern of stimuli that were removed during the training phase of the experiment. These were not presented to the participants. The colours of the stars is irrelevant.

*Category training*. In this phase, participants were trained on two thirds of the available stimuli. The training stimuli were selected randomly for each participant subject to several constraints: 1) that those stimuli selected were symmetrical around the category boundary and 2) that no adjacent stimuli of similar difficulty were removed (for example see Figure 4.1). In total there were 360 training trials, split into 3 blocks of 120 trials. In each block, the 24 stimuli were shown 5 times in a random order. On each trial, the participants looked at the stimulus until they made a response using either the “Z” key for Category A or the “/” key for Category B. Participants were unable to respond until at least 500ms had passed. Then, either “Correct” in green or “Incorrect!” in red was displayed for 500ms. A blank white screen was displayed between each trial for 500ms. Throughout the experiment, the labels “Category A” and “Category B” were displayed on the bottom left and right of the screen respectively. If participants took longer than 5000ms to respond, no corrective feedback was given, instead the message “PLEASE RESPOND FASTER” was displayed for 500ms.

*Recognition test*. In this phase participants judged whether each stimulus was “old” and appeared in the training phase, by pressing the “O” key, or was “new” and had not been shown in the training phase, by pressing the “N” key. The words “New” and “Old” were presented on the bottom left and right of the screen respectively. After this, participants

judged the confidence they had in their old-new judgement on a Likert scale that varied from 1 (=guessed) to 5 (=certain) by pressing the corresponding number key. Each of the 36 stimuli were presented three times in a randomised order. No feedback was given.

*Category test*. In this phase, participants were asked to judge the category membership of all 36 stimuli, not just those they had seen in the category training phase. No corrective feedback was given in this phase. Otherwise, the procedure was identical to that of the training phase. Each of the 36 stimuli were presented three times in a random order.

*Verbal report questionnaire*. Additionally, at the end of the experiment, participants were asked to complete a questionnaire that asked them to describe in detail the strategy that they used. This was to determine whether the participants could explicitly report the strategy they used and whether any participants used a rule-plus-exception strategy. The questionnaire asked them to "Imagine that another participant was asked to complete the experiment as you did. What instructions would you give them so that they could exactly copy your pattern of responding?" Participants were given a large blank box in which to write their answer.

**Analysis**

Additional strategy analyses are included in Section 4.5.

### 4.2.2 Results

The raw data will be archived at `www.willslab.co.uk/ply40` once this experiment is published. One outlier was removed from the rule-based condition as they scored below chance (0.5) on the categorisation task.

To examine recognition performance I calculated $d_a$ as the estimates of $d'$ varied across the confidence rating levels (Macmillan & Creelman, 2005). This is calculated by the following

$$d_a = \left(\frac{2}{1+s^2}\right)^{\frac{1}{2}} [z(H) - sz(F)] \tag{4.1}$$

where $s$ is the slope of the receiver operating characteristic (ROC) curve, $H$ is the hit rate and $F$ is the false alarm rate. I also calculated the adjusted response bias rate $c_a$ as

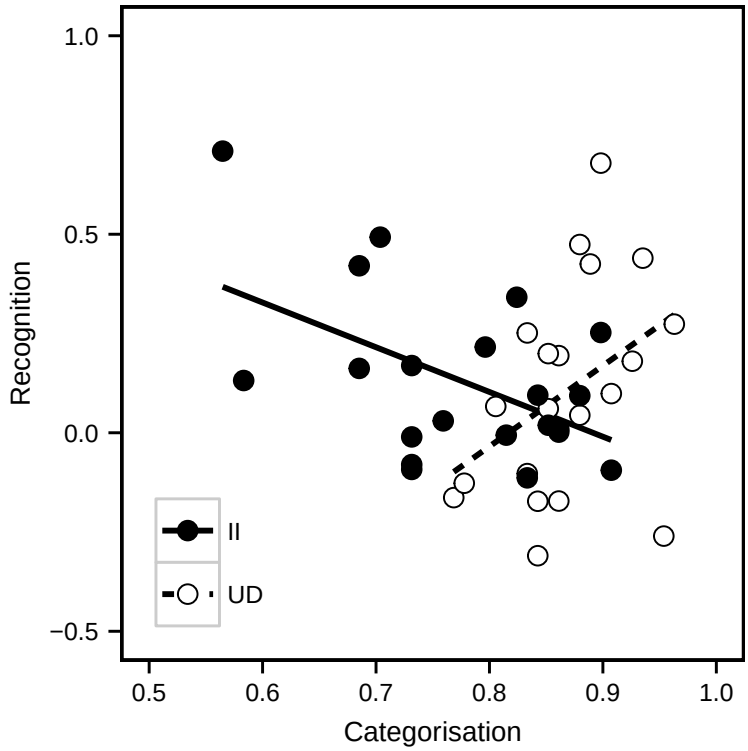


*Figure 4.2:* The relationship between categorisation accuracy and recognition in Experiment 7. Each point represents a participant, each line the line of best fit for each condition. Conditions: II=Information-integration, UD=Unidimensional.

follows

$$c_a = \left( \frac{-\sqrt{2}s}{(1+s^2)^{\frac{1}{2}}(1+s)} \right) [z(H) + z(F)] \tag{4.2}$$

A t-test found that recognition performance in the information-integration condition was significantly different from chance, $d_a = 0.13$, $SD = 0.21$, $t(20) = 2.80$, $p = .011$. Additionally, the difference between performance in the unidimensional condition and chance failed to reach significance, $d_a = 0.09$, $SD = 0.27$, $t(20) = 1.58$, $p = .129$. However, the difference between recognition memory performance between the unidimensional and information-integration category structure conditions failed to reach significance, $t(40) = 0.51$, $d = 0.16$, $p = .615$. There were also no significance differences in response bias between the unidimensional and information-integration category structure conditions, $t(40) = 1.00$, $d = 0.31$, $p = .323$.

It may also be informative to look at the relationship between categorisation and recognition performance. COVIS would predict that good performance on the information-integration category learning task is due to participants switching from the explicit, Verbal

System to the implicit, Procedural System. Therefore, it would predict that as performance on the information-integration categorisation task improves, recognition performance should decrease. Whereas, all participants in the rule-based categorisation task should have high recognition performance as they are using the Verbal System.

My predictions were based on the assumption that participants who learn the information-integration category structure well use complex, rule-based strategies. Therefore, one might expect that participants who perform better on the categorisation task use more complex strategies and so may also perform better on the recognition task. Whereas, good performance on the unidimensional categorisation task is unlikely to be correlated with recognition performance as all the participants will use the same, simple strategy. Therefore, we might expect that participants in the rule-based condition would all have poor recognition performance.

An ANCOVA found a significant interaction between category structure conditions and the relationship between category learning performance and recognition, $F(1,37) = 8.34$, $p = .004$. This interaction was driven by participants in the information-integration condition, $F(1,19) = 6.61$, $R^2 = 0.26$, $p = .019$. Recognition performance could be predicted from categorisation performance according the formula: recognition = 1.00 - 1.13*categorisation. For the rule-based condition this relationship did not reach significance, $F(1,18) = 3.43$, $R^2 = 0.16$, $p = .081$.

### 4.2.3 Discussion

The COVIS model predicts that categorisation is mediated by two competing learning systems that optimally learn different types of category structure. A key prediction of COVIS is that category knowledge acquired by the Procedural System will be unavailable to consciousness. In contrast, recent behavioural and neuroscientific work indicates that participants learning information-integration categories can access category knowledge and may be using explicit memory to facilitate categorisation (e.g. Chapters 2 and 3, Carpenter et al., 2016).

The current experiment aimed to directly test whether participants had access to knowledge about the information-integration category structure by comparing participants' performance on an old-new recognition task after learning either unidimensional rule-based or information-integration category structures. In this experiment, I found no evidence of superior memory for exemplars after learning a rule-based category structure compared

to learning an information-integration category structure. However, the relationship between categorisation and recognition performance depended on the category structure learned. For participants in the information-integration condition, as performance on the categorisation task improved, recognition performance fell. For participants in the rule-based condition, the trend was the other way around, although this did not reach statistical significance; as performance on the categorisation task improved so did recognition performance. Furthermore, contrary to the predictions of COVIS, recognition performance was not significantly above chance.

One possible concern with the results of Experiment 7 might be that the average recognition performance was low in both conditions. Therefore, in Experiment 8 I aim to improve average recognition memory performance by increasing the length of each trial.

## 4.3 Experiment 8

One reason I may have failed to find a clear recognition performance advantage for participants in either category structure may have been because of the quick pace of the experiment. Typically, in memory experiments participants see the stimuli and feedback for longer. For example, Sakamoto and Love (2004) displayed the stimulus and feedback for an extra 2501ms after the participant responded. It is possible that this helps raise recognition memory performance. This is especially important here as the stimuli are less perceptually discriminable as those in the Sakamoto and Love study. Therefore, in Experiment 8 the length of each element of the experiment protocol were lengthened.

### 4.3.1 Method

#### Participants

The participants were 44 undergraduate psychology students recruited from the Plymouth University participation pool. They were compensated with 2 research credits for their participation.

#### Stimuli and category structures

These were identical to the stimuli used in Experiment 7. The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness (12 participants) or size of the square stimuli (12 participants)–and two information-integration category structures–where the optimum boundary had either a positive (10

participants) or negative gradient (10 participants).

**Materials**

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

**Procedure**

The procedure for this experiment was identical to that of Experiment 7, apart from the fact that the timings were increased. In the category training and test phases, participants were unable to respond to the stimulus for 2000ms and the inter-trial interval was 1000ms. In the recognition phase, participants were not allowed to judge whether the stimuli were old or new for 2000ms. The confidence interval was also displayed for 2000ms before the participant could respond. The inter-trial interval remained 500ms.

**Analysis**

Additional strategy analyses are included in Section 4.5.

### 4.3.2 Results

The raw data will be archived at `www.willslab.co.uk/ply51` once this experiment is published. One participant from each condition were removed as they scored less than chance on the categorisation task.

As in Experiment 7, to examine recognition performance we calculated $d_a$ as the estimates of $d'$ varied across the confidence rating levels (Macmillan & Creelman, 2005).

A t-test failed to find a significant difference between recognition memory performance between the unidimensional and information-integration category structure conditions, $t(42) = 0.21,\ d = 0.06,\ p = .835$. Additionally, recognition performance was not found to be significantly different from chance for either the unidimensional, $d_a = -0.01$, $SD = 0.25$, $t(23) = -0.19$, $p = .848$, or information-integration conditions, $d_a = 0.01$, $SD = 0.25$, $t(19) = 0.11$, $p = .916$.

An ANCOVA failed to find any statistically significant relationships between conditions, category learning performance and recognition, $Fs < 1$. There were also no significant differences in response bias between the unidimensional and information-integration category structure conditions, $t(40) = 1.00$, $d = 0.31$, $p = .323$.

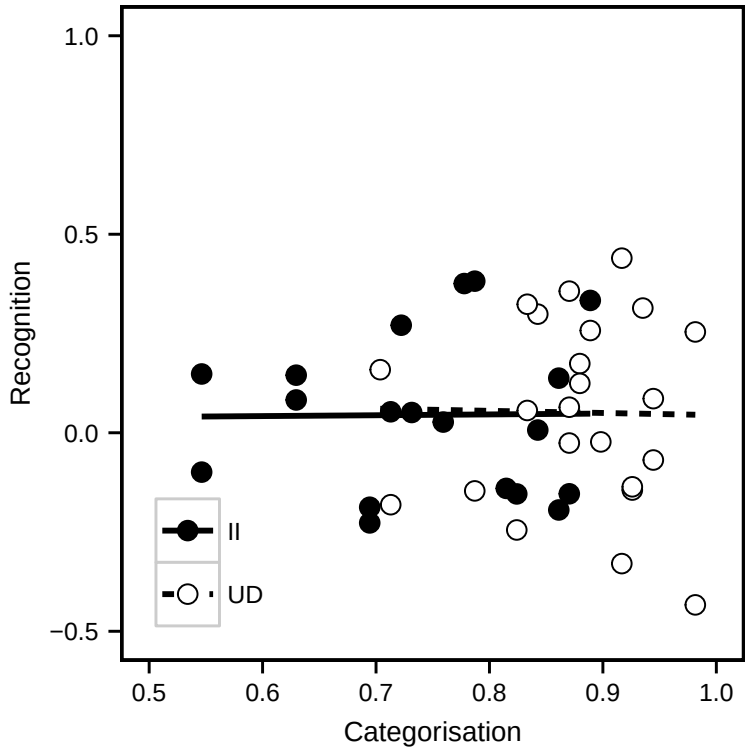


*Figure 4.3:* The relationship between categorisation accuracy and recognition in Experiment 8. Each point represents a participant, each line the line of best fit for each condition. Conditions: II=Information-integration, UD=Unidimensional.

### 4.3.3 Discussion

The literature that supports the COVIS model of category learning assumes that rule-based category structures are learned explicitly using the Verbal System and that information-integration category structure are learned implicitly using the Procedural System. However, recent fMRI data found that the medial temporal lobe was activated more in information-integration category tasks than rule-based tasks (Carpenter et al., 2016). Additionally, participants in the experiments reported in Chapters 2 and 3 were able to report the strategies they used in both category structures. In Experiment 7, I looked to see whether participants do actually learn information-integration tasks implicitly by examining subsequent recognition memory performance. I found an interaction between recognition memory, categorisation performance and category structure.

However, in Experiment 7 overall memory performance was poor. Therefore, in Experiment 8, I repeated Experiment 7 whilst giving participants more time to consider the stimuli and feedback. This was to evaluate the possibility that participants had poor recognition memory performance in both category structure conditions because they had not had enough time to process the stimuli sufficiently. This attempt to raise recognition per-

formance was unsuccessful as participants in both conditions did not have recognition performance significantly different from chance.

One reason Experiment 8 may have failed to find a difference between category structure conditions, is that information-integration category learning was poor. In this experiment, the majority of participants in the unidimensional category structure task performed much higher than those in the information-integration category structure task. For the rule-based unidimensional category structure, participants scored similarly in this experiment ($M = 0.88$, $SD = 0.07$) and in Experiment 7 ($M = 0.87$, $SD = 0.05$). However, this was not matched by a corresponding improvement in recognition memory for individual exemplars. Indeed, recognition memory performance was actually slightly worse in this experiment ($M = 0.05$, $SD = 0.24$) than the previous one ($M = 0.10$, $SD = 0.27$). Therefore, it appears that performance on the unidimensional task is unrelated to recognition memory for exemplars in that task.

For the information-integration category structure, categorisation performance in this experiment ($M = 0.75$, $SD = 0.11$) was slightly lower than in Experiment 7 ($M = 0.78$, $SD = 0.10$). Recognition performance for the current experiment ($M = 0.05$, $SD = 0.20$) was also slightly lower than the previous experiment ($M = 0.13$, $SD = 0.21$). Critically, here, average performance dropped below 75%, the maximum score a participant could get if they were using a unidimensional strategy to complete the information-integration category task. Therefore, it is possible that the majority of participants in both category structure conditions were using unidimensional strategies to complete the category learning task. However, I would only predict improved memory performance in the information-integration category structure task compared to the rule-based task, if participants were using complex, rule-based strategies. Similarly, COVIS would predict superior memory performance for the rule-based task compared to the information-integration task only if participants in the information-integration task were using the Procedural System. Again, this is unlikely considering the the average accuracy of these participants. Therefore, this experiment is not diagnostic of either COVIS or the rule-based approach I advocated.

One question remains: why does increasing the length of each trial have a differential effect on rule-based and information-integration category structure learning? One possibility is that by increasing the length of stimulus presentation time, participants would have more time to compare the image in front of them with other images in memory. This

may result in the stimuli becoming more confusable and thus reducing recognition accuracy. Therefore, in the following experiment I increased the perceptual discriminability of the stimulus.

This experiment increased the perceptual discriminability of the stimuli by adding a non-diagnostic stimulus dimension. Previous demonstrations of recognition memory in category learning used stimuli that varied on five binary dimensions (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). The stimuli used here vary on two continuous dimensions, size and brightness. This may mean that it is both harder to distinguish stimuli in the recognition phase from one another and to apply complex rules in the first place. Therefore, in Discussion I added a non-diagnostic stimulus dimension to the stimuli. This was a Fourier descriptor, which has been previously demonstrated to be difficult to verbalise (LaShell, 2010).

Additionally, participants in this experiment were trained to a learning criterion: participants had to reach 90% before the end of training. This was to ensure that participants learning the information-integration category knew the category structure as well as the participants in the other learning condition learned the unidimensional category structure. This would encourage participants to learn the information-integration category structure well: either using the Procedural System according to COVIS, or using complex rule-based strategies as I hypothesised.

### 4.3.4 Method

#### Participants

The participants were 62 psychology students recruited from the Plymouth University participation pool. They were compensated with 2 research credits for their participation.

#### Stimuli and category structures

The stimuli used in this experiment were cells formed of a cell-wall created from Fourier transforms and a nucleus of a grey circle (see Figure 4.4 for examples). The two relevant stimulus dimensions are the size and brightness of the nucleus.

The Fourier descriptors varied on the integral and non-commensurate dimensions of amplitude and phase (LaShell, 2010; Op de Beeck, Wagemans, & Vogels, 2003). The Fourier descriptors were constructed of three different sine-waves of frequencies of 2, 4 and 8 cycles per perimeter. The sine wave of 2 cycles per perimeter had amplitude of

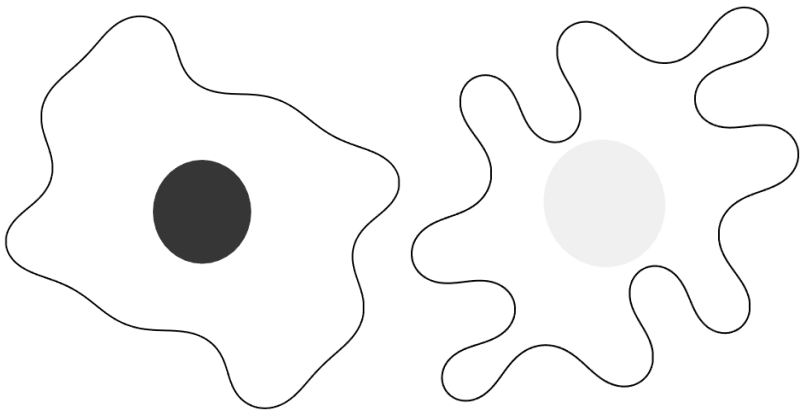

*Figure 4.4:* Examples of the stimuli used in Discussion.

0.25 radians and phase of 0 degrees. The sine wave of 4 cycles per perimeter had amplitude of 0.55 radians and phase of 0 degrees. The sine wave of 8 cycles per perimeter varied in amplitude and phase. Each dimension consisted of 6 possible values resulting in a total of 36 stimuli. The amplitude was between 0.5 and 1.4 radians and the phase was between 0 and 220 degrees.

For each participant, each circle stimulus was randomly assigned one of the Fourier descriptors. This random assignment was to make sure that the relevant stimulus dimensions were uncorrelated with the amplitude and phase of the Fourier descriptors.

The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness or size of the circle stimuli–and two information-integration category structures–where the optimum boundary had either a positive or negative gradient.

**Materials**

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

**Procedure**

The procedure for this experiment was identical to that in Experiment 7, aside from the fact that during the training phase participants were trained to criterion: participants were trained until they scored over 90% or reached a set number of blocks of training, whichever was sooner.

I found it difficult to set the number of total blocks that I expected participants to learn the information-integration category structure by. Originally, the maximum number of training blocks was set as 8. However, I found that although all the participants in the unidimensional category structure condition had reached the criterion by then, the majority (60%) of participants in the information-integration condition had not reached the criterion by the end of training. Therefore, I ran this condition again, the second time extending the maximum number of training blocks to 14. This resulted in 20 participants learning the unidimensional category structure and 26 learning the information-integration category structure (8 from the 8-maximum block version and 18 from the 14-maximum block version).

**Analysis**

All data analyses were conducted in R (R Core Team, 2015). The data analyses below are conducted on those who reached the learning criterion before the end of training.

### 4.3.5 Results

The raw data will be archived at `www.willslab.co.uk/ply76` once this experiment is published. One participant was removed from the information-integration category structure condition because they scored less than 0.5 at the category learning task at test, despite reaching the learning criterion during training.

A t-test failed to find a significant difference between recognition memory performance between the unidimensional and information-integration category structure conditions, $t(44) = 0.48$, $d = 0.14$, $p = .631$. However, recognition performance was found to be significantly different from chance for the information-integration condition, $d_a = 0.13$, $SD = 0.27$, $t(25) = 2.45$, $p = .022$, but not for the unidimensional condition, $d_a = 0.09$, $SD = 0.33$, $t(19) = 1.17$, $p = .255$.

Again, to examine recognition performance I calculated $d_a$ as the estimates of $d'$ varied across the confidence rating levels (Macmillan & Creelman, 2005). I also calculated the adjusted bias rate $c_a$.

An ANCOVA failed to find any statistically significant relationships between conditions, category learning performance and recognition, $Fs < 1$. There were also no significant differences in response bias between the unidimensional and information-integration category structure conditions, $t(44) = 0.50$, $d = 0.15$, $p = .622$.

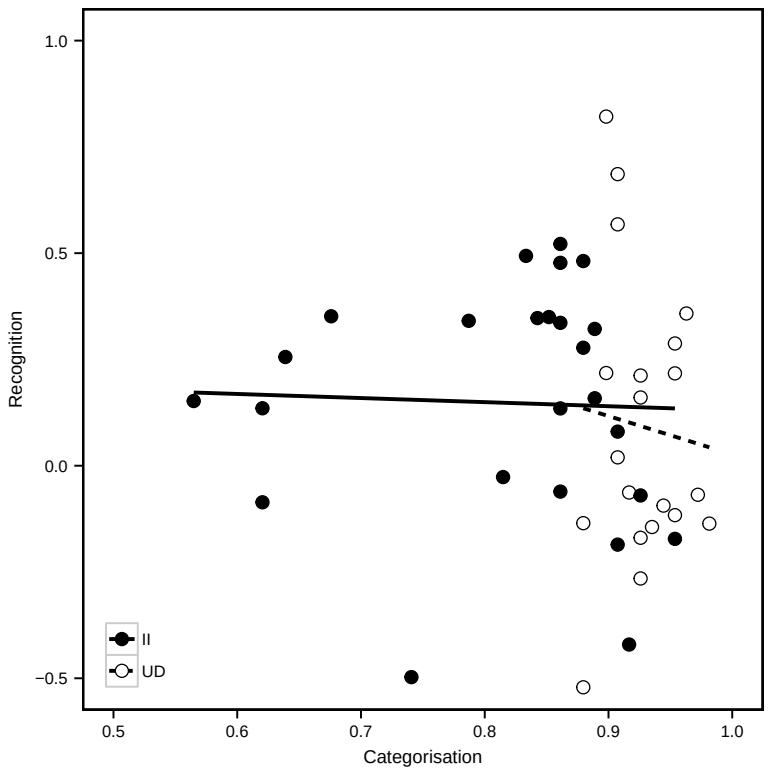


*Figure 4.5:* The relationship between categorisation accuracy and recognition in Discussion. Each point represents a participant, each line the line of best fit for each condition. Conditions: II=Information-integration, UD=Unidimensional.

### 4.3.6 Discussion

The COVIS model of category learning predicts that participants learning an information-integration category structure should have no explicit access to category knowledge, whereas participants in the rule-based category structure should.

In Experiments 7 and 8, I found slight evidence to suggest the contrary. Participants in the information-integration category structure condition appear to have slightly superior recognition memory compared to those who learned a rule-based category structure. The current experiment found similar evidence once participants had been trained to criterion. As in Experiment 7, recognition performance was significantly higher than chance in the information-integration condition but not in the unidimensional condition.

One issue with the design of Discussion is that as the information-integration category structure is more difficult to learn, participants in this condition took more blocks to get to criterion than those in the rule-based condition. In other words, participants in the information-integration conditions on average saw the stimuli more times than those in the rule-based conditions. To address this issue, in Experiment 10 I extended the training so that all the participants saw the stimuli the same number of times but participants in

the more difficult information-integration condition had enough time to learn.

## 4.4 Experiment 10

### 4.4.1 Method

#### Participants

Forty participants were recruited from the Plymouth University participation pool. They were rewarded with 4 points (partial course credit). Additionally, the participant that scored the highest in each counterbalance condition during training was awarded a £15 Amazon voucher.

#### Stimuli and category structures

The stimuli were generated in the same way as Discussion although the Fourier transform borders were not included.

The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness (10 participants) or size of the circle stimuli (10 participants)–and two information-integration category structures–where the optimum boundary had either a positive (10 participants) or negative gradient (10 participants).

#### Procedure

The procedure for this experiment was identical to that for Discussion aside from all participants were trained for 20 blocks on the categorisation task.

#### Analysis

All data analyses were conducted in R (R Core Team, 2015).

### 4.4.2 Results

A t-test failed to find a significant difference between recognition memory performance between the unidimensional and information-integration category structure conditions, $t(38) = 0.90$, $d = 0.28$, $p = .354$. Additionally, recognition performance was not found to be significantly different from chance for either the unidimensional, $d_a = -0.01$, $SD = 0.18$, $t(19) = -0.32$, $p = .753$, or information-integration, $d_a = 0.05$, $SD = 0.24$, $t(19) = 0.88$, $p = .389$, conditions.

Again, to examine recognition performance I calculated $d_a$ as the estimates of $d'$ varied

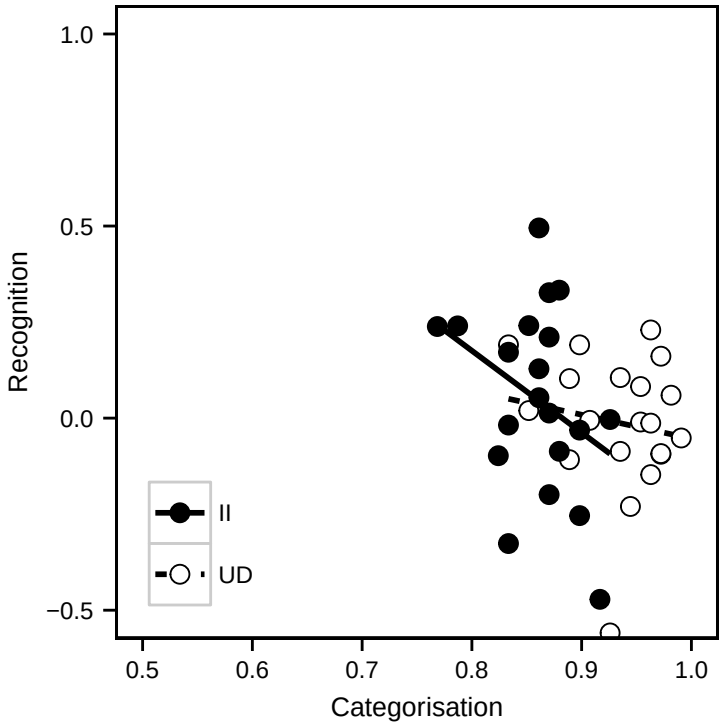


*Figure 4.6:* The relationship between categorisation accuracy and recognition in Experiment 10. Each point represents a participant, each line the line of best fit for each condition. Conditions: II=Information-integration, UD=Unidimensional.

across the confidence rating levels. I also calculated the adjusted bias rate $c_a$.

An ANCOVA failed to find any statistically significant relationships between conditions, category learning performance and recognition (see Figure 4.6): the test of the relationship between categorisation and accuracy approached significance, $F(1,36) = 3.13$, $p = .085$. For the remaining comparisons $Fs < 1$.

There were also no significant differences in response bias between the unidimensional and information-integration category structure conditions, $t(38) = 1.28$, $d = 0.40$, $p = .209$.

### 4.4.3 Discussion

The COVIS model of category learning predicts that participants learning an information-integration category structure should have no explicit access to category knowledge, whereas participants in the rule-based category structure should. To test this theory, I conducted four experiments to see whether participants had superior recognition memory performance after learning a unidimensional rule-based category structure compared to an information-integration one. I found that this was not the case: in three out of four experiments, there were no significant differences in recognition performance between category structure conditions. This appears inconsistent with the predictions of the CO-

VIS model.

That being said, the predictions of the COVIS model are also contingent on the strategies that participants use. COVIS predicts that the two learning systems implement different strategy types which are compatible with different types of category structure. However, the experimental evidence uses this in the opposite direction: the category structure is experimentally manipulated with the hope that this will encourage participants to use the learning system that would be most compatible. In the COVIS literature, researchers used a GRT-informed model-based strategy analysis to determine whether participants are using the optimum strategy in each condition (Maddox & Ashby, 1993). Therefore, to facilitate comparison with these studies, the model-based strategy analysis is also included below.

## 4.5 Strategy analyses

### 4.5.1 Model-based strategy analysis

The version of the strategy analysis used here is identical to that reported in Section 3.6.1 in Chapter 3. Briefly, here I fitted four strategy types to category responses given by each participant in the test phase of the experiment: unidimensional, diagonal (GLC), conjunction and random. The best model was selected for each participant using the BIC (Schwarz, 1978). Also reported are the average Schwarz weights for the winning models (Wagenmakers & Farrell, 2004).

As required by the COVIS model, the majority of participants in each condition are using the optimum strategy for the category structure they learned. In other words, in the rule-based condition, more participants are using the unidimensional strategy than any of the other strategies. Similarly, in the information-integration condition more participants are using the diagonal (GLC) strategy than any of the others. This analysis (as used by the proponents of COVIS) indicates that the category structure manipulation resulted in a corresponding switch in learning system.

That being said, I have serious reservations about the validity of this analysis. For instance, in Chapters 2 and 3, across 5 experiments, the strategy analysis corresponded very poorly with the strategies that the participants described using. This issue will be discussed in greater detail in Chapter 5. For the present, the take away message is that these experiments meet the criteria typically used in the COVIS literature and there-

*Table 4.1:* The proportion of participants in each experiment of Chapter 4 that were identified as using each strategy type using the model-based analysis.

| Category structure | Strategies *(wBIC)* | | | |
|---|---|---|---|---|
| | UD | GLC | CJ | RND |
| Experiment 7 | | | | |
| Unidimensional | 0.81 *(0.78)* | 0.10 *(0.98)* | 0.05 *(0.58)* | 0.05 *(0.69)* |
| Information-integration | 0.29 *(0.80)* | 0.57 *(0.97)* | 0.10 *(0.65)* | 0.05 *(0.44)* |
| Experiment 8 | | | | |
| Unidimensional | 0.83 *(0.82)* | 0.12 *(0.84)* | 0.04 *(0.67)* | - |
| Information-integration | 0.20 *(0.72)* | 0.50 *(0.93)* | 0.15 *(0.66)* | 0.15 *(0.66)* |
| Discussion | | | | |
| Unidimensional | 0.75 *(0.81)* | 0.05 *(0.54)* | 0.20 *(0.75)* | - |
| Information-integration | 0.04 *(0.57)* | 0.92 *(0.91)* | 0.04 *(0.76)* | - |
| Experiment 10 | | | | |
| Unidimensional | 0.80 *(0.79)* | 0.15 *(0.74)* | 0.05 *(1.00)* | - |
| Information-integration | - | 0.90 *(0.95)* | 0.10 *(0.86)* | - |

Strategies: UD=Unidimensional, GLC=General linear classifier, CJ=Conjunction, RND=Random.

fore, the results of these experiments can be generalised to other experiments within this literature.

### 4.5.2 Verbal report analysis

In the experiments in the previous chapter, it was also enlightening to examine the categorisation strategies that participants reported using. In the previous chapters, these varied considerably from the strategies identified by the model-based strategy analysis ubiquitously used in the COVIS literature (Maddox & Ashby, 1993). COVIS would predict that the majority of participants in the rule-based category structure condition would report a rule-based category structure, whereas most of those in the information-integration category structure condition should not be able to describe the strategy they used. In contrast, I would predict that participants in all conditions would report using rule-based strategies, but that they would be more complex in the information-integration condition than the unidimensional condition.

The verbal reports were independently coded by the author (CERE) and a volunteer (AJW in Experiment 7, GW in the remaining experiments). First, each verbal report was examined to determine whether the participant had reported an explicit categorisation strategy or not. The inter-rater reliabilities for this in each experiment are reported in

*Table 4.2:* Summary of inter-rater reliability tests for each experiments in Chapter 4.

| Experiment | Explicit strategy | | Strategy Type | | Removed | |
|---|---|---|---|---|---|---|
| | $\kappa$ | $p$-value | $\kappa$ | $p$-value | UD | II |
| Experiment 7 | 1.00 | <.001 | 1.00 | <.001 | 2 | 4 |
| Experiment 8 | 0.85 | <.001 | 0.77 | <.001 | 3 | - |
| Discussion | 1.00 | <.001 | 0.70 | <.001 | 2 | 3 |
| Experiment 10 | 1.00 | <.001 | 0.90 | .001 | 2 | 3 |

Table 4.2. The several participants in each experiment that failed to report a strategy reported extraneous details about the experimental setup such as listing the response keys or describing the different phases of the experiment. Table 4.2 shows how many of these participants were excluded from the strategy tables below according to experiment and category structure condition. Finally, the type of categorisation strategy was identified. The inter-rater reliability for strategy assignment is also reported in Table 4.2.

In addition to the strategies already identified in Chapters 2 and 3, here we also included an *other* category. This was for people who reported using a rule based on a stimulus dimension not the brightness and size of the stimuli. Typically, these strategies referred to the order of the stimuli in some way, such as “If the stimulus is darker than the previous stimulus it is in Category A.”

In Table 4.3, we can see that the majority of participants described using rule-based strategies. In all the experiments, over 80% of the participants in the unidimensional category structure conditions reported using the optimal unidimensional strategy. For the information-integration category structure conditions, the majority of participants reported using rule-based strategies using both stimulus dimensions, i.e. either a conjunction or a more complex strategy. This is consistent with the experiments reported in Chapters 2 and 3. It is also consistent with my prediction that participants use complex rule-based strategies to learn information-integration category structures.

## 4.6 Bayesian analyses

The strategies described by participants reveal a limitation of the analyses comparing recognition performance between category structure conditions. These reports show that the variation of strategy types between participants is higher in the information-integration category structure than in the rule-based unidimensional one. When learning the unidi-

*Table 4.3:* The proportion of participants that reported using each strategy for the experiments in Chapter 4.

| Category structure | Strategies | | | | |
|---|---|---|---|---|---|
| | UD | Implicit | CJ | Complex | Other |
| Experiment 7 | | | | | |
| Unidimensional | 0.84 | 0.05 | 0.05 | 0.05 | - |
| Information-integration | 0.22 | - | 0.56 | 0.17 | 0.06 |
| Experiment 8 | | | | | |
| Unidimensional | 0.90 | - | - | 0.10 | - |
| Information-integration | 0.19 | - | 0.12 | 0.69 | - |
| Discussion | | | | | |
| Unidimensional | 0.84 | - | 0.05 | 0.11 | - |
| Information-integration | - | - | 0.04 | 0.96 | - |
| Experiment 10 | | | | | |
| Unidimensional | 0.95 | - | - | 0.05 | - |
| Information-integration | 0.05 | - | 0.58 | 0.16 | 0.21 |

Strategies: UD=Unidimensional, CJ=Conjunction.

mensional category structure, over 80% of participants use the optimal unidimensional strategy. However, when learning the information-integration category, participants use a wide range of two dimensional rule-based strategies. Additionally, in some experiments, a non-trivial subset of participants in this condition report using either a unidimensional strategy or another idiosyncratic strategy (such as stimulus order). Arguably, these type of strategies could be regarded as less optimal than the two dimensional strategies, as they demonstrate that the participant was not aware that the optimal boundary depended on both stimulus dimensions.

This difference in strategy type variability is important because the predictions I made were contingent on the strategies that participants used to complete the categorisation task, not the type of categorisation task itself. In other words, the unidimensional condition appears to be a good marker for unidimensional responding, but that the information-integration condition appears be a poor marker for complex rule-based responding. Therefore, it may be that participants using complex rules have much greater recognition performance than participants using simpler strategies, but this effect is obscured by averaging over a disparate group of participants. To address this possibility, in this section I will use Bayesian methods to compare recognition performance between participants who used a single dimension to classify stimuli with those who used multiple dimensions.

Bayesian methods are the ideal tool here for several reasons. First, as mentioned previously, Bayesian techniques will allow us to judge whether there is truly no difference in recognition performance between strategy types or whether the data were just not diagnostic (Dienes, 2011). Second, these techniques can be used to combine data from several experiments in a principled way (Kruschke, 2015). This attribute is especially helpful for the work presented here as difference in recognition performance between strategy types is predicted to be small. This is because, although a complex two-dimensional rule-based strategy would require participants to attend to both stimulus dimensions, this may only result in a slight improvement in recognition performance. By combining the data from multiple experiments, the sensitivity of the analysis is improved thereby improving the likelihood that I may be able to detect a small but consistent effect. Third, using Bayesian techniques avoids some of the pitfalls of using null hypothesis significance testing (Dienes, 2011; Kruschke, 2013). For example, $p$-values are not only unintuitive and difficult to understand, but also only incidentally answer the question: what does the data say about the likelihood of my hypothesis being true (Dienes, 2011; Kruschke, 2015).

The Bayesian techniques in this section use Markov chain Monte Carlo (MCMC) algorithms to estimate the size of the effect of strategy type on recognition memory performance (as measured by $d_a$; Kruschke, 2013, 2015). The procedure here closely follows that of Kruschke (2013) and the reader should turn there for a detailed explanation as only the main points are reviewed here.

The first step of this analysis is to specify a descriptive model for the data from all four experiments reported here. In this case, I will be using the $t$-distribution

$$d_a \sim t(\mu_i, \sigma_i, \nu) \tag{4.3}$$

to describe the distribution of recognition memory scores. A $t$-distribution has three parameters: the mean $\mu$, the standard deviation $\sigma$ and the shape parameter $\nu$ (more generally known as the degrees of freedom). When $\nu$ is small the distribution has heavy tails, and when $\nu$ is large the distribution tends to the normal distribution. This means that the distribution can easily describe data both with and without outliers (Kruschke, 2013). Each strategy type group (unidimensional and multi-dimensional) will be defined by a $t$-distribution, with different means and standard deviations. The shape parameter will be the same for both groups. Therefore, the descriptive model of recognition will have

5 parameters: two mean values ($\mu_1$ and $\mu_2$), two standard deviations ($\sigma_1$ and $\sigma_2$) and a single shape parameter ($\nu$).

The second step is to determine the prior distribution (Kruschke, 2013). This describes the distribution of credibility across parameter values. In other words, which values of the parameters described above are most likely. To reduce subjective bias, it is important that the prior is broad and is defensible to a skeptical audience.[‡] The priors on the mean parameters $\mu_1$ and $\mu_2$ were assumed to be a broad normal distribution. The mean of this prior distribution was arbitrarily set to the mean of the pooled data. To keep the distribution broad, the standard deviation was set as 1000 times the standard deviation of the pooled data. The priors on the standard deviation parameters $\sigma_1$ and $\sigma_2$ were assumed to be uniformly distributed from a low value $L$ of one thousandth of the standard deviation of the pooled data to a high value $H$ 1000 times the standard deviation of the pooled data. The prior on the shape parameter $\nu$ was exponentially distributed so as to balance evenly the credibility of a normal and non-normal shape. The priors described here are vague, so the prior will have little impact on the estimation of any effect and will let the data speak for itself.

The third step is to use MCMC algorithms to reallocate credibility over the parameter values so it is more consistent with the data. Practically, this is achieved using Bayes' rule (Equation 4.6; Bayes & Price, 1763). Bayes' rule calculates the probability of the parameters given the data (posterior) from the probability of the data given the parameter values (likelihood) and the probability of the parameter values (prior). These are divided by the probability of the data (evidence).

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \tag{4.4}$$

Calculating the value of the evidence in this case is impossible to do analytically. Instead, we rely on numerical methods that avoid ever having to explicitly evaluate it. The posterior distribution is estimated by sampling from it using MCMC algorithms. To do this, I have used models and code written by Kruschke (2013) implemented in R (R Core Team, 2015). This also uses the MCMC sampling language JAGS, available in R using the

[‡]It can be informed by previous evidence. However, as there are relatively few studies that have looked at recognition memory in category learning and none using this particular procedure, for these experiments I followed Kruschke (2013) in using a relatively uninformative prior.

package rjags (Plummer, 2003).

The data used for this simulation were combined from Experiments 7-10. Any participants that reported using either an “Implicit" or “Other" strategy were excluded, as well as any participants who failed to report a strategy. This resulted in 77 participants who reported using each strategy type. The credible parameter values were generated by sampling from the posterior distributions 100,000 times.

Interpreting the posterior distribution can be done in two ways: by looking at the estimated values of the parameters or looking at the highest density intervals (HDI) of the parameter estimates (Kruschke, 2013). The HDI summarises the distribution by an interval that includes the points of the distribution that are most credible and cover most of the interval. It is defined as the interval that spans 95% of the interval for which all of the points in the interval have a higher probability density than those outside the interval.

This method of Bayesian estimation mirrored the findings of the null hypothesis testing for each experiment above. Modal recognition performance when using unidimensional rules $\mu_1$ was estimated at 0.044. However, the HDI ranged from -0.15 to 0.104 indicating that it is still credible that recognition performance did not differ from chance. The modal standard deviation $\sigma_1$ was estimated as 0.1 with an HDI between 0.209 and 0.304.

The modal recognition performance when using complex rules $\mu_2$ was estimated at 0.1, with a HDI between 0.046 and 0.149. This indicates that all of the credible values for recognition memory for participants using complex rules were above chance. The modal standard deviation $\sigma_2$ was estimated as 0.218, with a HDI between 0.181 and 0.263. The modal logged shape parameter was estimated as 1.44, with an HDI between 0.835 and 2.080.

The modal difference in recognition performance between complex and unidimensional strategies was estimated as 0.051, with an HDI between -0.24 and 0.134 (positive numbers indicate superior memory with complex rule-based strategies than simple unidimensional strategies).

In Bayesian estimation, a hypothesis is tested against the null hypothesis by comparing the relevant posterior distribution with a region of practical equivalence around the null (ROPE; Kruschke, 2013, 2015). The ROPE encloses the values that are thought to be negligibly different from zero and thus, the same as the null hypothesis for all practical purposes. Here, I have defined the ROPE as between -0.05 and 0.05 as the effect I

would expect to see is very small (Cohen, 1998). The null hypothesis is accepted if the majority of credible values lie inside the ROPE. In the current analysis, 9.8% of the credible values lie inside the ROPE, 4.7% of the values lie in favour of the predictions of COVIS and 85.5% of the values lie in favour of the prediction that there is superior recognition memory for participants who use complex rules compared to those using unidimensional rules

## 4.7 General discussion

The COVIS model describes category learning as mediated by two independent learning systems: one explicit, the other implicit (Ashby et al., 1998, 2011). Furthermore, these systems are hypothesised to optimally learn rule-based and information-integration category structures respectively. However, despite these claims, no studies have directly examined whether participants are learning information-integration category structures implicitly. Rather proponents of COVIS have focused on examining the procedural nature of learning in the implicit Procedural System (Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2016). From this, they have inferred that learning is implicit as demonstrations of procedural learning in other domains have been found to be implicit (such as in Willingham et al., 2000).

In contrast, I found indirect evidence that participants are learning information-integration category structures explicitly. In the experiments reported in Chapters 2 and 3, I asked a large number of participants to report the strategy they had used to complete the learning task. The majority of participants were able to report the strategy they had used, even when learning an information-integration category structure. Furthermore, Carpenter et al. (2016) found greater activation in the medial temporal cortex in information-integration category structure learning than in rule-based learning. As the medial temporal lobe has long been considered critical for explicit memory processes (Conroy et al., 2005; Squire, 1992), this suggests that information-integration category learning involves explicit memory processes to a greater extent than rule-based category learning.

In this chapter I aimed to directly test whether information-integration category structures are learned implicitly as predicted by the COVIS model. To do this, participants learned either a rule-based or information-integration category structure. They were then given an old-new recognition memory test. Recognition memory is argued to be an explicit process (Berry et al., 2012; Gabrieli & Fleischman, 1995). I found no evidence to sup-

port the predictions of COVIS. Across four experiments, there was not any evidence that there was superior recognition memory for exemplars in rule-based category structures than information-integration category structures. Furthermore, the majority of participants were able to report the strategy they used. Additionally, when looking at the strategies that participants report using in the information-integration conditions, the majority of participants reported using complex, explicit, rule-based strategies based on both stimulus dimensions.

Additionally, Bayesian estimation techniques allowed me to compare memory performance between those participants who reported using unidimensional and those using complex two-dimensional rules. This analysis supported the prediction that participants would have superior recognition memory performance when using complex rules compared to using simpler rules.

However, the results of the ANCOVA analyses seem to add slightly more complexity to the story. More specifically, for participants in the unidimensional conditions, there is little evidence of a relationship between categorisation accuracy and recognition performance. For the information-integration category structure conditions, the relationship seems to vary. Experiments 8 and 9 did not find a relationship between categorisation accuracy and recognition. However, in Experiments 7 and 10 there appeared to be a negative correlation between categorisation accuracy and recognition performance: as categorisation accuracy improved recognition accuracy decreased. Therefore, it may not just be the type of strategy that participants are using that increases recognition memory, but also how well the strategy performs. If a participant is aware that their strategy is not working well, they may make more effort to remember past exemplars in order to try and infer a strategy that better accounts for the data.

Although speculative, this hypothesis is consistent with previous evidence that investigated the links between categorisation and recognition memory. These studies found that participants have greater recognition memory for exceptions to a simple rule compared to a control stimulus (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). These studies demonstrate that participants are able to recall single stimuli as well as implement simple rules and indicate that category learning may not be as simple as either exemplars or rules, but a combination of both. To examine the possibility that both strategy type and strategy performance have a role in participants' memory for exemplars, a future exper-

iment might attempt to vary the error rate of certain stimuli regardless of the strategy or responses given by the participant. If error rates are important, we may expect to see that participants with a falsely elevated error rate may have superior memory than those without. Additionally, those who report using complex strategies might also have superior memory than those using simpler ones.

### 4.7.1 Limitation

One limitation to the experiments reported above is that the strategies that participants used are only vaguely defined by the participants' own verbal reports. This means that it is incredibly difficult to predict which exemplars a particular participant might remember better. As I would predict superior recognition memory for only a few key exemplars in the information-integration condition, this means the effect I am searching for is incredibly small and based on only a few exemplars. Therefore, the results presented here are less clear than one might hope.

A solution to this problem might be to define the strategies that participants use with more precision. One way of doing this might be to use the model-based analysis used by proponents of COVIS to precisely define decision boundaries through space for each participant. Theoretically, this analysis would allow me to identify the key "exception" stimuli and then compare them with control stimuli in the other condition as in previous work (Palmeri & Nosofsky, 1995). However, as mentioned previously, the model-based strategy analysis used in the COVIS literature assigns participants to strategy types that do not correspond to the strategies that participants report using. Most interesting is the observation that participants who report using complex rule-based strategies are most often assigned by this analysis as using the diagonal (GLC) strategy, which is hypothesised as being a marker for implicit processing and the Procedural System. This issue is examined much more fully in Chapter 5.

Another interesting thing to note is the differences in category learning performance between Chapter 3 and the experiments reported here. The information-integration category structure used in this chapter is conceptually similar to that used in Chapter 3. The only difference being that 6 extra stimuli were added to simplify removing a third of the stimuli in the training phase to be assigned as new. This relatively minor change has had a large impact on participants ability to learn the category structure. By the end of training in the experiments reported in Chapter 3, participants were scoring around 90%.

But in this chapter, participants were scoring from between 70% to 90%, sometimes after significantly more training.

### 4.7.2 Conclusion

In summary, the experiments presented here indicate that participants using complex rule-based strategies have superior recognition memory than those using unidimensional strategies. Furthermore, the majority of the participants use these complex strategies to learn information-integration category structures. These findings undermine the ubiquitous assumption in the COVIS literature that participants learn information-integration category structures implicitly.

# Chapter 5

# Strategy analysis

In Chapters 2, 3 and 4 I reported a number of experiments where the majority of participants who learned an information-integration category structure were identified by the GRT-informed model-based analysis as using a diagonal (GLC) strategy. In the COVIS literature, using this optimum diagonal strategy to solve an information-integration task is taken as evidence that participants were responding using the Procedural System, thus responding implicitly. However, the strategies identified by this analysis contrast sharply with those participants reported using; rather they reported using complex, two-dimensional, rule-based strategies. These verbal reports cast doubt on the assumption that the diagonal (GLC) strategy model is a marker of implicit responding. In this chapter, I report the results of several simulations that resolve this apparent contradiction.

## 5.1 Introduction

Usually in the study of category learning, researchers focus on how learning performance varies between groups rather than looking at individual differences (Estes, 1956; Maddox, 1999). In typical categorisation experiments, each participant is presented with a series of stimuli and asked to sort them, one by one, into categories (Levering & Kurtz, 2015). Often these experiments compare learning performance between participants of different category structures under different conditions (in other words, a between-subjects 2x2 design). For example, in Ell, Ing, and Maddox (2009) participants learned one of two category structures and received feedback either immediately after making a response or after a delay. Then, to draw conclusions about the mechanisms of category learning from these types of experiments, researchers compare the average accuracy performance in each condition. In the case of Ell et al., they inferred that feedback delay only affects learning of some types of category structure. To make these inferences, researchers must assume that all the participants in a particular condition learn the same way (Estes, 1956). Thus, by averaging over groups of participants, relatively little information is lost

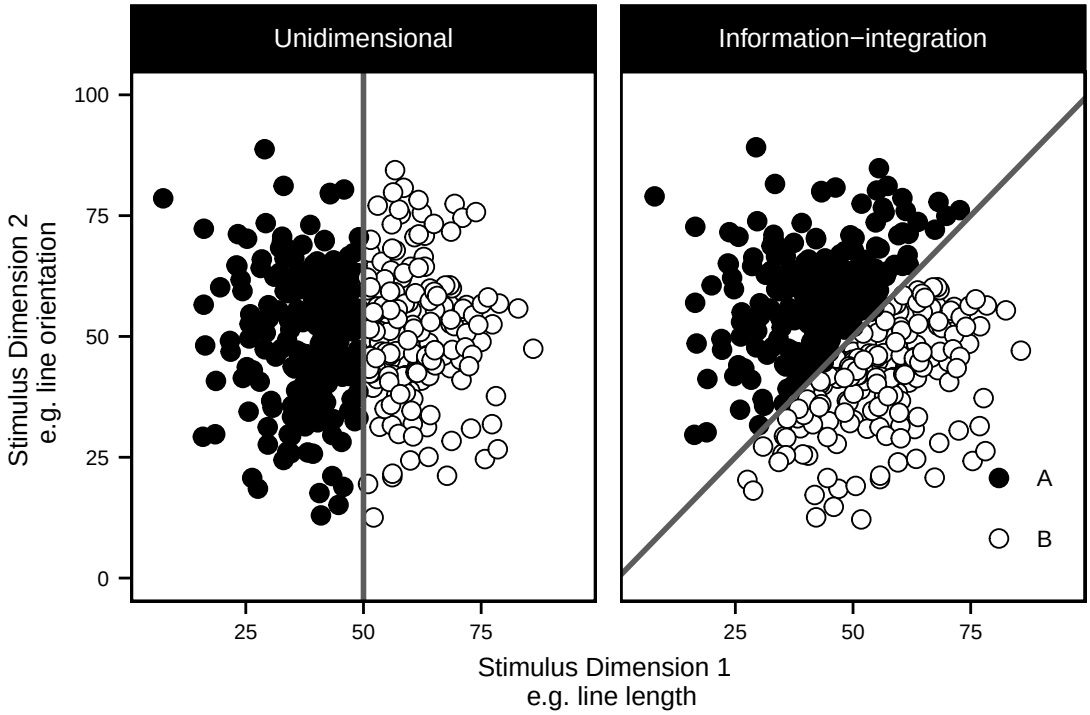


*Figure 5.1:* Two example strategies implemented as linear decision boundaries through a hypothetical two-dimensional stimulus space. Each point represents a stimulus.

and valid conclusions can be made.

However, there is compelling evidence that, for any categorisation task, there will be groups of participants who complete the task differently, despite having identical training histories (e.g Maddox & Ashby, 1993; Meeter, Myers, Shohamy, Hopkins, & Gluck, 2006; Nosofsky & Zaki, 2002; Wills et al., 2015). In other words, for any condition of an experiment, there will be subsets of participants who will give different consistent patterns of behaviour in response to the stimuli. One of these subsets will probably be those using the optimum strategy for the category structure they are learning. However, other subsets of participants will use strategies that do not result in optimum performance. For example, some participants may categorise stimuli based on only one stimulus dimension (as in Figure 5.1a), even if optimum performance on the task requires using multiple stimulus dimensions (as in Figure 5.2b). Indeed, some suggest that it is typical for participants to begin categorisation tasks by applying simple rules, only moving to more complicated response strategies later if they are motivated to do so (e.g. Ashby et al., 1998; Milton et al., 2008; Raijmakers, Dolan, & Molenaar, 2001; Wills et al., 2013).

The consequence of participants using qualitatively different strategies in a category learning task is that the inferences made from averaged performance data are no longer

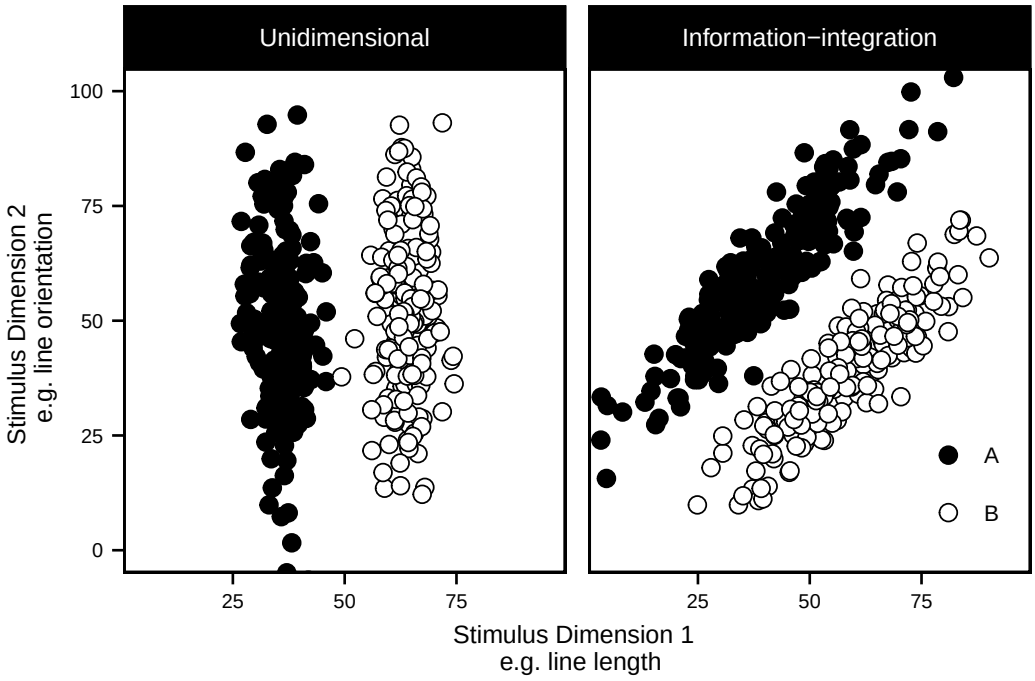


*Figure 5.2:* Two example category structures: a) unidimensional rule-based, b) information-integration (Smith et al., 2015).

valid (Estes, 1956; Maddox, 1999). To illustrate this point, consider an experiment that compared the effect of a manipulation, such as concurrent load, that results in a reduction of performance by 10% for a particular category structure (such as in Zeithamova & Maddox, 2006). In the ideal case, all participants would be using the same optimum strategy and all those in the relevant condition would be similarly affected by the manipulation; participants with concurrent load would score 10% less than they would have without the load. In this case, we could carry on using our standard performance analyses. However, if some participants are using other, sub-optimum strategies then the interpretation of the experiment is less clear. One alternative possibility is that the manipulation changes the proportion of strategies used in each condition. This would result in a change in average accuracy because, given a particular category structure, the highest level of performance possible for each strategy varies. Another possibility is that the manipulation did not cause a change of strategies, but might have a differential effect depending on the strategy type being used. For example, the manipulation could have had no effect on people using the optimum strategy, but could severely interrupt performance on the sub-optimum strategies (see Schnyer et al., 2009, for a similar argument).

The COVIS model goes part way to describing why and when participants may use different strategies (Ashby et al., 1998, 2011). It hypothesises that the Verbal System im-

plements rule-based strategies (such as see Figure 5.1a). Because the Verbal System implements simple rule-based strategies, COVIS predicts that this system will optimally learn category structures that are implementations of simple rules, such as the unidimensional structure shown in Figure 5.2. If rule-based strategies do not result in high enough accuracy–because the category structure is not rule-based and thus difficult to verbalise–COVIS predicts that the Procedural System will gain control of responding. As the Procedural System can implement a variety of strategies (including the one demonstrated in 5.1b) it can implement the optimum strategy for the information-integration category structure shown in Figure 5.2.

That being said, the experiments used to support the COVIS model reverse this logic; the researchers manipulate category structures between participants in order to elicit a switch in the category learning system the participants use (see Chapter 1 for a review). These researchers predict that because participants are learning a rule-based or information-integration category structure they will learn to use the appropriate strategy and therefore be using the Verbal or Procedural System, respectively. This can be problematic because these experiments rely heavily on dissociation logic (Newell et al., 2010). Typically, they compare learning of both rule-based and information-integration category structures under the influence of another factor (see Ashby & Maddox, 2005, 2011, for reviews). If this factor affects learning of one category structure more than another, they infer that the factor affects one system more than the other. Critically, these types of dissociation are predicated on the assumption that the majority of participants used the optimum system to learn each category structure. If this is not the case, any differences in overall accuracy between category structure conditions might be due to different rates of sub-optimum strategies between conditions, rather than to the existence of two category learning systems.

To avoid the possibility that any dissociation in accuracy is due to different proportions of sub-optimum strategies (and thus sub-optimum learning systems), proponents of COVIS use a strategy analysis informed by GRT (Ashby & Gott, 1988; Ashby & Soto, 2015) to determine which strategy each participant is using (Maddox & Ashby, 1993). Then, they compare the strategy each participant is using with the category structure they were assigned to learn. If a sufficient number of participants are found to be using the optimum strategy for the category structure they are learning, then the category structure manipulation is assumed to have elicited a corresponding shift in category learning system. If

a particular experiment meets this criterion, then any dissociations in accuracy can be validly ascribed to the existence of dual-systems, bar confounds.

This chain of inferences puts the analyses that identify participants' strategies in a critical position in determining whether or not an experiment supports COVIS. If for some reason the strategy analysis was invalid, then an experiment previously thought to support COVIS would actually indicate a differential effect of the manipulations on the strategies that participants use. For instance, consider an experiment that found that feedback delay harmed information-integration category learning but had no effect on unidimensional rule-based category learning (such as Ell et al., 2009). Furthermore, suppose that the strategy analysis found that all the participants used the optimum strategy for the category structure they learned. If the strategy analysis was accurate, we could conclude that the source of this interaction was due to the properties of two different systems. However, if the strategy analysis was inaccurate this inference would not be the only one we could make. For example, if the strategy analysis falsely identified a conjunction rule as a diagonal strategy in the information-integration conditions, a more parsimonious account might be that feedback delay impacts learning once participants are using sufficiently complex strategies. Therefore, it is crucial that the analysis used to identify these strategies is valid and reliable. In the following section, I will review the key features of the strategy analysis used in the COVIS literature and in the experiments in the previous chapters. I will then evaluate whether this analysis meets the standards required for its role in the COVIS literature.

### 5.1.1 Strategy analysis details

Recall that the strategy analysis approach used in the COVIS literature aims to find, for each participant, the decision bound through stimulus space that best separates their responses for one category from the other*. This approach is inspired by GRT, a multidimensional extension of signal detection theory (Ashby & Gott, 1988; Ashby & Soto, 2015). Different strategy types are implemented by varying the functional form of the decision boundary. For example, a unidimensional strategy is implemented by a straight line perpendicular to one of the stimulus dimensions. This approach also assumes that stimulus perception is subject to normally distributed noise: each time a participant sees

*For ease of explanation we have focused on the case where the stimuli under consideration have two stimulus dimensions and are being sorted into two categories. This is not always the case, although it is the most frequent instantiation of this theory and analysis within the COVIS literature.

a particular stimulus it is perceived slightly differently. Therefore, stimuli near the decision boundary are more likely to be misclassified as small amounts of noise may make them appear to the participant as if that stimulus was on the wrong side of the boundary. Each participant's strategy is determined by fitting multiple decision bound models, with different functional forms, to their response data and some measure of fit is calculated for each model. The functional form of the model that best describes that participant's pattern of responding gives their strategy.

The two kinds of model-based strategies that are of particular interest within the COVIS literature are rule-based and information-integration strategies (Ashby & Gott, 1988; Maddox & Ashby, 1993). This is because these kinds of strategy type are hypothesised to be implemented by the Verbal and Procedural Systems of COVIS respectively (Ashby et al., 1998, 2011). Within the COVIS literature, rule-based strategy models are implemented by linear decision boundaries that are parallel or perpendicular to the stimulus dimensions in stimulus space (Maddox & Ashby, 1993). Typical examples include unidimensional and conjunction rules. A unidimensional rule consists of a single decision boundary orthogonal to the relevant dimension. It has up to two parameters: perceptual variance and the value of the boundary on the relevant dimension. For example, a unidimensional rule on basis of line length corresponds to the rule "If the line is short it is in Category A, if it is long it's in Category B." A conjunction rule consists of two decision boundaries orthogonal to each other. It has up to four parameters: perceptual variance in two dimensions and the values where the two boundaries cross the axes. For example, this might correspond to the rule "If the line is short and upright it's in Category A; otherwise Category B." These strategies are assumed to be implemented by the Verbal System and to be optimum for rule-based category structures.

The information-integration strategy model is usually implemented in the COVIS literature by a diagonal strategy, also known as the General Linear Classifier (GLC; Maddox & Ashby, 1993). The diagonal strategy consists of a single linear decision boundary in stimulus space which is not parallel to any of the stimulus dimensions. This strategy has up to three parameters: perceptual variance, the slope and y-intercept of the line. This strategy is generally difficult to verbalise and so is hypothesised to be implemented by the Procedural System and be the optimum strategy for learning information-integration category structures (Ashby et al., 1998, 2011).

In addition to the sets of models that correspond to the Verbal and Procedural Systems of COVIS, researchers within the COVIS literature also include models of non-learning. These random models do not assume the existence of a decision boundary. Rather they assume that stimulus features are irrelevant to responding and that participants respond at random. There are two types of random model usually included: one with no parameters that assumes that participants respond equally to both categories, and one with one parameter which represents biased responding towards one category.

Once the best fitting strategy model has been selected from the models above for each participant using a measure of fit, proponents of COVIS use these strategies to check the experimental assumptions of COVIS. Typically, they look at the proportion of each strategy type in each cell of the experimental design. For any dissociation in accuracy scores to support the COVIS model of category learning, there must be a corresponding difference in the types of strategy identified between the category structure conditions. Specifically, there must be more rule-based strategies than other strategies in the rule-based category structure conditions, and more diagonal (GLC) strategies than other strategies in the information-integration category structure conditions. If this is the case, researchers assume that their category structure manipulation was successful in inducing an equivalent change in category learning system. Then, any changes in participants' learning of each category structure can be attributed to the manipulation differentially affecting the underlying learning system mechanism.

## 5.2 Strategy analysis validity

Using the GRT-informed strategy analysis as a manipulation check for the experimental evidence for COVIS is logically valid—as long as it both consistently and accurately identifies the response strategy that participants are actually using. In other words, the analysis must be able to correctly identify a participant's strategy under a variety of circumstances (and levels of noise). For the COVIS literature, what matters the most is that the strategy analysis can correctly identify participants' strategies when they are learning a variety of category structures. If the strategy analysis performed well for rule-based category structures but poorly for information-integration category structures, it would not be possible to infer whether there was a change in strategies between category structure conditions because of a change in learning system or because of the strategy analysis.

Arguably the most common category structures used in the COVIS literature are the uni-

dimensional and information-integration category structures generated by the randomisation technique proposed by Ashby and Gott (1988). Examples are shown in Figure 5.2. In these structures, each category $i$ is generated by sampling points from a bivariate normal distribution with mean $\mu_i$ and covariance matrix $\Sigma_i$. Each point represents a stimulus, with the $x$-value corresponding to one stimulus dimension and the $y$-value corresponding to the other stimulus dimension. It is common for researchers to first generate the stimuli for the unidimensional category structure and then rotate the points $\pi/4$ radians around the average of both category distributions (the 'centre of gravity' of the points) to get the information-integration category structures. These category structures are argued to be excellent choices for comparing explicit and implicit category learning because they differ markedly in verbalisability whilst being matched on several key attributes such as within-category similarity, between-category distance and the optimal accuracy a participant could achieve (usually $95\%$ or above; Smith et al., 2014, 2015). In the following, I consider the effectiveness of the strategy analysis in relation to the unidimensional and information-integration category structures separately.

### 5.2.1 Information-integration category structure

In the experiments reported in the previous chapters, the largest divergence between the verbal reports given by participants and those identified by the model-based analysis was found for participants who learned the information-integration category structure. The majority of verbal reports described using rule-based strategies that used information about both stimulus dimensions. In contrast, the model-based analysis tended to identify these participants as using the diagonal (GLC) strategy, the optimum strategy for the information-integration category structure.

One possible explanation is that the model-based analysis is biased by the category structure towards finding the optimal strategy for that structure. Indeed, Donkin et al. (2015) found preliminary evidence for this when they compared the strategies identified by the analysis on the basis of stimuli that formed an information-integration category structure with those identified when additional transfer stimuli were added to cover the whole space. They found that the proportion of information-integration strategies identified were higher for the clearly information-integration category structure than when the transfer stimuli were included. Therefore, it is possible that the diagonal (GLC) strategy would be more likely to be identified for participants who learned an information-integration cate-

*Table 5.1:* The proportion of each type of strategy recovered for each type of generating strategy for an information-integration category structure.

| Generating strategy | Recovered strategies *(wBIC)* | | | |
|---|---|---|---|---|
| | UD | CJ | GLC | RND |
| UD | 0.99 *(0.93)* | - | - | 0.01 *(0.06)* |
| CJ | 0.09 *(0.32)* | 0.45 *(0.01)* | 0.46 *(0.58)* | - |
| GLC | 0.04 *(0.29)* | 0.01 *(0.03)* | 0.87 *(0.94)* | 0.07 *(0.23)* |

Strategies: UD=Unidimensional, CJ=Conjunction, GLC=General linear classifier, RND=Random.

gory structure, regardless of the strategies participants were actually using.

**The simulation**

To test this intuition, I conducted a model-recovery analysis to examine the accuracy of the model-based strategy analysis in discriminating between these strategies. This model-recovery procedure is recommended as best practice for any cognitive modelling analyses (Heathcote, Brown, & Wagenmakers, 2014). It involves simulating hypothetical participants' responses according to the strategy models actually used by the strategy analysis. From these hypothetical, simulated participants I can then use the model-based strategy analysis to identify the strategies from the responses to see whether it is capable of recovering the correct generating model. The advantage of simulating different response strategies, rather than looking at real participants who took part in experiments, is that I can be absolutely certain of what the input to the analysis is and so can fairly evaluate the model-based analysis.

The details of the model-recovery procedure are briefly described here. For the full details see Appendix A. First I generated an information-integration category structure using the procedure outlined by Ashby and Gott (1988) according to the parameters reported by Smith et al. (2015). Then, I determined the optimum strategy for that category structure according to either unidimensional, conjunction or diagonal (GLC) strategies. I then simulated the responses of 20 hypothetical participants according to this optimum strategy. To these responses, I added noise. Finally, I conducted the model-based analysis on each participant and calculated the proportion of participants who were identified as using the unidimensional, conjunction, diagonal (GLC) and random strategies. The results from this simulation are shown in Table 5.1. Also reported are the Schwarz weights for each model (Wagenmakers & Farrell, 2004).

The optimum diagonal (GLC) strategy is identified well, as is the unidimensional strategy. In these cases, any misinterpretation that the strategy analysis made would not have any theoretical impact as the generating strategy is either correctly identified or identified as the theoretically neutral random strategy.

In contrast, the mis-identification of the conjunction strategy is seriously problematic for the behavioural evidence for COVIS. Participants whose responses were generated using a conjunction strategy are equally likely to be identified as using a diagonal (GLC) strategy as the correct conjunction strategy. In other words, of the simulated participants using an optimal CJ strategy, 46% were incorrectly identified as using a GLC strategy. Therefore, the existing literature likely has a larger proportion of participants using a rule-based strategy to learn an information-integration category structure than estimated by the strategy analysis. Possibly this proportion is even larger than estimated here as this simulation included only one type of two-dimensional rule-based strategy whereas the participants in the previous chapters could report many more complex strategies.

The consequence of this high proportion of mis-identified participants is that we are likely to accept false evidence that supports the existence of two learning systems (Type I error). This type of mis-identification would make it appear that the category structure manipulation was successful when it was not: participants were still using rule-based strategies to learn an information-integration structure. Then, researchers might incorrectly infer that a manipulation designed to impair implicit learning had successfully done so. However, the drop in performance would more likely be due to participants using more complex, rule-based strategies in the information-integration category structure condition than in the rule-based one.

### 5.2.2 Unidimensional category structure

So what about the unidimensional category structure? When considering the validity of the strategy analysis applied to a unidimensional category structure, it is first necessary to distinguish between the strategy that participants are using and the pattern of responses given. COVIS makes predictions about the strategies participants are using. In other words, it makes predictions about how the participant represents the strategy they use to sort the stimuli into two categories. The strategy analysis aims to determine this representation, but the data it uses to do this are the responses the participants gives to individual stimuli. The analysis cannot evaluate why a participant assigned a particular

stimulus to a particular strategy, only that they did. In some respects, this objectivity could be a strength as it avoids some of the problems associated with introspection. However, it also means that it is critical that the stimuli presented to the participant would allow the analysis to discriminate between the different possible strategy representations.

Unfortunately, with the unidimensional category structure typically used (e.g. Smith et al., 2015) it is incredibly difficult to discriminate between different strategy representations. This is because it is possible to achieve 100% accuracy using strategies other than the optimal unidimensional strategy. For example, a conjunction strategy can also score 100% if the vertical bound passes between the two category structure and the horizontal bound passes either above or below one of the categories. A diagonal (GLC) strategy can also score 100% if the line is steep enough so as to pass between the two categories. Furthermore, it seems likely that the strategy analysis will identify all of these possible strategies as unidimensional ones. This is because the model-based strategy analysis uses measures of model fit that penalise the number of parameters in each model (Akaike, 1974; Schwarz, 1978). Therefore, if a participant used a diagonal (GLC) strategy on a unidimensional category structure, it is more likely that the strategy analysis would identify that participant as using the optimum unidimensional strategy as the unidimensional strategy has two parameters whereas the diagonal (GLC) strategy has three.

If this were the case, it would have serious implications for the COVIS literature. COVIS predicts that participants should be using a verbalisable rule-based strategy when learning a unidimensional strategy. However, if there is no way of determining from the model-based analysis whether a participant has been identified as using a unidimensional responder because they are actually using that strategy. Therefore, when framed in terms of the COVIS model, the unidimensional category structure cannot discriminate between the Verbal and Procedural Systems.

**The simulation**

To test this intuition I conducted a simulation similar to the one in the previous section (for more details see Appendix A). First I generated a unidimensional category structure using the procedure outlined by Ashby and Gott (1988), similar to that used in Smith et al. (2015). Then, I fit a unidimensional, conjunction and diagonal (GLC) strategy to this category structure to determine the optimum strategy of that type for this category

*Table 5.2:* The proportion of each type of strategy recovered for each type of strategy.

| Generating strategy | Recovered strategies *(wBIC)* | | | |
|---|---|---|---|---|
| | UD | CJ | GLC | RND |
| UD | 0.96 *(0.95)* | 0.01 *(0.00)* | - | 0.03 *(0.16)* |
| CJ | 0.82 *(0.88)* | 0.09 *(0.05)* | 0.08 *(0.37)* | 0.01 *(0.08)* |
| GLC | 0.52 *(0.50)* | 0.17 *(0.00)* | 0.25 *(0.41)* | 0.06 *(0.20)* |

Strategies: UD=Unidimensional, CJ=Conjunction, GLC=General linear classifier, RND=Random.

structure. For example, the optimum diagonal (GLC) strategy for this category structure is a steep diagonal line that passes between the two categories.† I then simulated the responses of 20 hypothetical participants according to this optimum strategy. To these responses, I added additional noise. Finally, I conducted the model-based analysis on each participant and calculated the proportion of participants who were identified as using the unidimensional, conjunction, diagonal (GLC) and random strategies. The results from this simulation are shown in Table 5.2. Also reported are the Schwarz weights for each model (Wagenmakers & Farrell, 2004).

This simulation confirms my intuition: the GRT-informed strategy analysis cannot discriminate well between strategies applied to the unidimensional category structure. Note first, that the analysis does well when the generating strategy is unidimensional (the first row of Table 5.2). Here, the recovered strategy models are either the correct unidimensional strategy or the random model. That some hypothetical unidimensional responders are identified as using a random strategy is not problematic as the random model is a-theoretical within the COVIS literature. For COVIS, the random model is hypothesised to be those participants who did not learn the category structure. Therefore, unidimensional generating strategy models recovered as random strategies does not result in false inferences, such as participants using the incorrect learning system.

Overall, the model-based strategy model does well in recovering the unidimensional strategy models. However, if we take this as a marker of good performance, we can see that the more complex models are recovered much less well. For example, 9% of the simulated participants who were using a conjunction strategy were correctly identified,

†Of course, analytically a vertical line is a special case of the diagonal (GLC) strategy. However, in practice, the model fitting procedure still attempts to estimate a gradient for the line. Then, because the model is fitted using an algorithm that uses gradient descent, there comes a point were even a large increase in the gradient of the line results in little to no improvement in fit. Thus, in practice, the optimum diagonal (GLC) model is close to vertical, but still diagonal.

whereas 82% of them were mis-identified as using a unidimensional strategy. Similarly, only 25% of the simulated participants who used the diagonal (GLC) strategy were correctly identified. Instead, 69% of them were found to be using a rule-based strategy (either unidimensional or conjunction strategies). Also, interesting are the values for the Schwarz weights. For the diagonal generating strategy, the Schwarz weights would indicate that you should trust the recovered unidimensional strategies more than the correct diagonal strategy.

For COVIS, misidentifying the diagonal (GLC) strategy is much more problematic than misidentifying the conjunction strategies in this way. This is because the diagonal strategy is assumed to be a marker for the implicit Procedural System and the unidimensional strategy is assumed to be a marker for the explicit Verbal System. If we assume this, then participants could be learning a unidimensional category structure implicitly, using the Procedural System, however, the analysis would misleadingly demonstrate that they were using the Verbal System instead.

The simulation demonstrates that the unidimensional category structure is not useful for identifying the strategies that participants are using. This adds to other literature that indicates that using unidimensional category structures may not be a good example of a rule-based category structure for the COVIS literature. For example, in Chapter 2 I argued that the number of stimulus dimensions required to optimally learn the category structure can confound dissociations between rule-based and information-integration category structures. Additionally, this analysis warns against using data from participants learning a unidimensional strategy to support evidence of participants using an explicit rule-based strategy.

One possible objection to this conclusion is that it is contingent on participants using the optimum version of each strategy type. In other words, the version of each strategy that scores the highest for this unidimensional category structure. In actual experiments, participants may be using the wrong type of strategy in a sub-optimal way. For example, using a diagonal (GLC) strategy whose decision boundary passes through both categories in a unidimensional category structure. Surely then, the strategy analysis would be able to better discriminate between strategy types? This is indeed a possibility. However, it seems reasonably unlikely to happen in practice due to the extremely high accuracy levels of participants in unidimensional category structures. For example, Ashby et al.

(2002) found an average accuracy of above 90% for participants in their unidimensional category structure conditions (see also Smith et al., 2015). Therefore, although this may be true in some situations, it is unlikely to be true in the majority.

Some may argue that, although theoretically the unidimensional category structure cannot discriminate between the strategies (or learning system) that participants may be using, in practice this has limited impact. After all, the majority of evidence indicates that participants begin training by using simple rules, based on one stimulus dimension (Raijmakers et al., 2001; Wills et al., 2015). It seems that participants only switch to more complex rules, using multiple dimensions, if they have the time, cognitive resources or inclination (Milton et al., 2008; Wills et al., 2013). Therefore, it is unlikely that participants would be use anything but a unidimensional strategy to learn the unidimensional category structure. Indeed, even if a few participants used a more complex strategy that was necessary, falling to identify them would have a relatively small impact.

## 5.3 Practical demonstration

The previous two simulations demonstrate that the model-based strategy analysis often misidentifies the strategies simulated participants use when learning either the unidimensional or information-integration category structures. Of particular concern for the COVIS literature is that rule-based and diagonal (GLC) strategies can be confused. If these strategy types were being misidentified in real experiments, it would undermine using the model-based strategy analysis as a manipulation check and cast doubt on all the experiments that rely on that logic.

However, these simulations lack one critical feature common to all real-world studies of category learning: accuracy scores. Typical categorisation studies focus on gross measures of performance and only use strategy analyses as a secondary measure, if at all (Levering & Kurtz, 2015). This makes it tempting to assume that strategy confusions would rarely happen in reality. In other words, because these sub-optimum strategies by definition cannot score as well as the optimum strategies, it would be difficult to achieve the average performance scores seen in the literature whilst misidentifying participants' strategies.

Some may argue that the participants using a sub-optimum strategy to complete a categorisation task would score so poorly that they would be removed by applying a learning criteria. In the category learning literature, participants who performed poorly (usually

under a certain percentage) are often excluded from any statistical analyses. This is because researchers are typically interested in the process of learning. If a participant scored poorly they likely did not learn so are of little theoretical interest (Newell et al., 2010). However, in the COVIS literature this is unlikely for two reasons. First, it is relatively rare for studies in this literature to use learning criteria. Indeed, in one case, not including a learning criterion resulted in the researchers misinterpreting their data and incorrectly claiming that their results supported dual-systems of category learning (Newell et al., 2010). Furthermore, if the authors of these studies do use learning criteria, they tend to be highly unsystematic. For example, in Smith et al. (2014) participants were excluded if "they showed significantly lower performance during their last 100 trials than during their first 100 trials" (p. 450). However, despite sharing the same authors as the Smith et al. (2014) paper, in Smith et al. (2015) participants were excluded if they scored below 70% in the final training block. Not only does varying the learning criterion make comparing results between experiments difficult, neither of these learning criteria are guaranteed to exclude participants using a sub-optimum strategy.

The second reason participants using sub-optimum strategies are likely to be included in analysis is that, for the category structures used in this literature, most sub-optimum strategies still result in performance significantly above chance. For example, using a unidimensional strategy (illustrated in Figure 5.1) in an information-integration category structure (such as in Figure 5.2) can result in participants scoring around 75%. This is much higher than typical learning criteria (for reference Newell et al., 2010, used a learning criterion of 65%), so any participants using a sub-optimum strategy would not be excluded.

In this section, I will aim to demonstrate that the model-based analysis will misidentify participants' strategies whilst still performing at levels of accuracy reported in published work. In the following, I demonstrate that it is possible for all participants to be using rule-based strategies but still find a) an interaction between an experimental manipulation and the learning of rule-based and information-integration category structures, and b) that the majority of participants will be identified as using the optimum strategy for each category structure by the model-based strategy analysis. The experiment I consider is Experiment 2 of Smith et al. (2015).

### 5.3.1 Experiment 2, Smith et al. (2015)

In this experiment, Smith et al. (2015) investigated the effects of time pressure on category learning. The experiment was split into a training and a test phase. In the training phase, participants either learned a unidimensional or an information-integration category structure (as in Figure 5.2) with feedback but without any time pressure. In the test phase, participants were then "tested" (they still received feedback) on the category structure they had learned under unspeeded and speeded conditions (the order of which was counterbalanced between participants). In the unspeeded test phase, participants had as much time to give their responses as they liked, whereas in the speeded test phase participants only had 600ms to respond. Smith et al. found an interaction between category structure and response deadline conditions. Participants learning with a response deadline generally performed worse, however this difference was greater for participants learning the information-integration category structure than those learning the rule-based, unidimensional category structure (for a similar result, see Milton et al., 2008).

Smith et al. (2015) is interesting to re-examine because of the assumptions they make. Smith et al. assume *a priori* that a) "dissociable category-learning utilities exist" (p.2478) and that b) explicit and implicit processes are accessed using rule-based and information-integration category structures. As this experiment is pitched as investigating the properties of the Verbal and Procedural System of COVIS, it follows that COVIS is unfalsifiable in this experiment: any result would be consistent with a dual-system approach because a dual-system model is already assumed.

Smith et al.'s (2015) overconfidence in a dual-systems account of category learning also results in Smith et al. failing to use the model-based analysis as a manipulation check. Instead, they use it as a secondary dependent variable to answer the question "does time pressure change the systems participants use to learn categories?" Smith et al. argued that the time pressure manipulation pushed participants in the information-integration condition towards using sub-optimum (i.e. unidimensional) strategies. These conclusions would be substantially undermined if the strategy analysis failed to correctly identify the strategies participants used.

Additionally, the explanation of Experiment 2 of Smith et al. seems inconsistent with other arguments within the COVIS literature. In most investigations of the processes of cate-

gory learning, concurrent load and time pressure are assumed to have similar effects (Wills et al., 2015). However, in the COVIS literature, concurrent load is predicted to harm rule-based category learning and not information-integration learning (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). This is the opposite prediction to that made by Smith et al. about time pressure.

Furthermore, Experiment 2 of Smith et al. (2015) could also be argued to be consistent with a single system approach. The evidence from the experiments in Chapters 2, 3 and 4 indicate that participants learn information-integration category structures using complex, verbalisable rules. For example in Chapter 2, I looked at the effect of feedback type on learning rule-based and information-integration category structures initially investigated by Ashby et al. (2002). As part of these experiments, the participants were asked to describe the strategies they used to categorise the stimuli. These verbal reports were then compared to the strategy the participants were assigned using the model-based analysis. As would be predicted by COVIS, the strategy analysis used there found more information-integration responders in the information-integration category structure conditions than in the rule-based category structure conditions. However, the verbal reports of all the participants identified as information-integration strategy responders by the model-based analysis described using a rule-based strategy. I failed to find any participants who could have possibly been describing using an implicit strategy, even when using very broad selection criteria. This was also found in Chapters 3 and 4: the strategies that participants report using, do not match well with those the strategy analysis predicts.

These experiments also indicate that the rules participants use are not identified by the model-based strategy analysis as rule-based strategies, rather as diagonal (GLC) strategies. Therefore, it is possible that the majority of participants in the information-integration category structure conditions could actually be using a conjunction or other two-dimensional rule-based strategy to respond in the unspeeded condition. However, when under time pressure they fall back on using a simpler rule-based approach. Consistent with this hypothesis, Wills et al. (2015) found that participants under time-pressure were more likely to use a sub-optimum rule-based approach, rather than switching to overall similarity sorting strategy.

This possibility is of particular import when considering Experiment 2 of Smith et al.

(2015) because the authors failed to include complex rule-based strategies as possibilities in their strategy analysis. This is a problem because Donkin et al. (2015) demonstrated that not including complex rules increased the number of participants that were misidentified as diagonal responders. Donkin et al. compared a model-based strategy analysis typical of experimental work within the COVIS literature with an extended strategy analysis that included a greater variety of both rule-based and information-integration strategy models. They found that when they included a larger selection of possible category learning strategies, the proportion of participants identified as using the optimum information-integration strategy fell. In other words, they found that some participants who were identified as using the optimum information-integration strategy in the restricted analysis, were found to be using a non-standard, more complicated, rule-based strategy in the extended analysis. This indicates that participants learning an information-integration category structure may only be identified as using an implicit diagonal structure because a sufficient number of rule-based models were not included in the model-based analysis. In other words, 'implicit' responders in the COVIS literature may actually be using complex, rule-based strategies.

Choosing the set of models to consider is difficult for researchers who use model-based approaches (although see Raijmakers, Jansen, & van der Maas, 2004, for an example of an approach not subject to this limitation). The crux of the problem is that there may always be a model that better accounts for the data than set of models included. Arguably, the extent of this problem depends on the relationship the model-based analysis has with the research question. For example, if you are trying to find out whether Model A is better than Model B at describing this data, then the existence of hypothetical other models is relatively unimportant. After all, in this research question you are just trying to find the best out of two possible models, not the best overall (although this may be your ultimate aim).

However, when a model-based analysis is used as a manipulation check, the possible existence of better fitting models is a much greater problem. In the case of the strategy analysis used in the COVIS literature, it may be that participants identified as "implicit" responders might be using complex, rule-based strategies that were not included in the set of possible strategy models. This result would appear to confirm that the category structure manipulation was successful; the category structure manipulation induced a corresponding change in the strategies participants used. However, participants in both

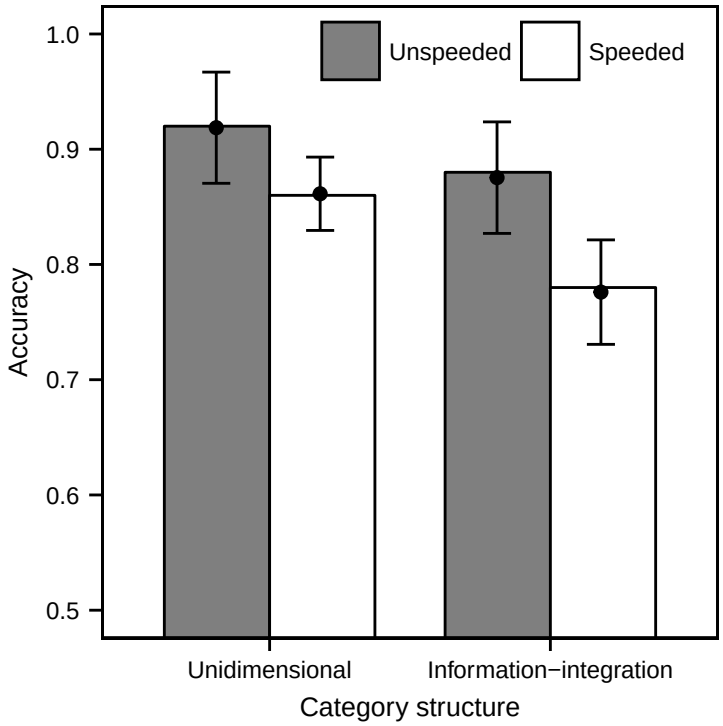


*Figure 5.3:* Graph representing the simulation of Smith et al. (2015) using sub-optimum strategies. Bars represent the accuracy scores reported by Smith et al. (2015) and points the accuracy achievable with participants using sub-optimum strategies.

rule-based and information-integration category structure conditions would both be using rule-based approaches, meaning that the category structure manipulation was not successful. This would result in a Type I error. Any differential effect of another independent variable on learning of these category structures could not be ascribed to the existence of two competing systems of category learning. It should rather be ascribed to the differential effect of the manipulation on different types of rule-based strategies.

### 5.3.2 The simulation

To see whether it was possible that all the participants in Experiment 2 of Smith et al. (2015) were using rule-based strategies, I first generated a set of hypothetical participants. These participants' responses were generated from unidimensional and conjunction strategy models that could best fit either the unidimensional or information-integration category structures used by Smith et al.. I then added various levels of noise to these hypothetical participants and calculated their accuracy. Then I performed the model-based analysis which included three model types: unidimensional, diagonal and random models. Note that although some simulated participants' responses were generated by a conjunction strategy, this strategy type was not included in the model-based analysis.

*Table 5.3:* The number of participants from Smith et al. (2015) and the simulation above that were assigned to each strategy type according to the model-based analysis for each condition.

| Condition | Strategies | | |
|---|---|---|---|
| | UD | GLC | RND |
| UD-Unspeeded | | | |
| Smith et al. (2015) | 16 | 14 | - |
| Simulation | 30 | - | - |
| UD-Speeded | | | |
| Smith et al. (2015) | 22 | 8 | - |
| Simulation | 30 | - | - |
| II-Unspeeded | | | |
| Smith et al. (2015) | 3 | 27 | - |
| Simulation | 3 | 27 | - |
| II-Speeded | | | |
| Smith et al. (2015) | 13 | 16 | 1 |
| Simulation | 13 | 16 | 1 |

Strategies: GLC=General linear classifier, UD=Unidimensional, RND=Random.

This was to keep the model-based analysis as similar to the one conducted by Smith et al. as possible. I then selected 30 hypothetical participants for each condition such that a) they had the same average accuracy as that reported in Experiment 2 of Smith et al. (see Figure 5.3) and b) were identified by the strategy analysis as using the same strategy types as reported in Figure 8 of Smith et al. (p. 2486; see Table 5.3) as much as possible.

In addition to these hypothetical participants having the same average accuracy and identified as using the correct strategies, it was also possible to replicate the statistical tests. Here, I found a main effect of category structure, $F(1,58) = 17.34$, $p < .001$, a main effect of time pressure, $F(1,58) = 10.43$, $p = .002$ and the critical interaction, $F(1,58) = 4.44$, $p = .037$.

Table 5.3 compares the strategies identified by Smith et al. (2015) with those identified by the simulation. Remember that in the simulation, *all* of the generating models were rule-based strategies, i.e. unidimensional or conjunction strategies. The most obvious difference is that in the simulation, no participants in the unidimensional category structure conditions were identified as using a diagonal (GLC) strategy. This is because all of the simulated participants who were identified by the analysis as using a diagonal (GLC) strategy had a total performance of less than 70% for the unidimensional category struc-

ture and so were excluded by the learning criterion. For the remaining strategies, the unidimensional strategy is a special case of the diagonal (GLC) strategy with less parameters. As these simulations used BIC to discriminate between strategies, and this criterion takes into account the number of parameters, the unidimensional strategy always win.

So why is this not the case in the study by Smith et al. (2015)? There are several possibilities. One possibility is that, by the process of learning, participants may have changed their representation of the stimuli. Studies have shown that ratings of within-category similarity increase and between-category similarity decrease across training. Therefore, it is possible that this process has warped the stimulus space so that unidimensional strategies appear to be diagonal (GLC). However, as this process has not occurred in our simulations, the unidimensional strategy still wins. Another possibility is that participants are using a sub-optimum version of a conjunction strategy in the unidimensional category structure, which is more easily confusable with a unidimensional strategy.

A final possibility is due to how I applied the learning criterion. Smith et al. (2015) applied the 70% learning criterion to the training data. However, here I assumed that participants were excluded based on their final test performance. This may actually be a more stringent criterion as there is some evidence that participants in the test phase of Smith et al. were scoring well below 70%. For example, in the information-integration deadline condition, one participant was found to be using a unidimensional strategy whose decision boundary crossed the $x$-axis at around 18 (Figure 8; Smith et al., 2015). This strategy would likely score closer to 50% as nearly all the stimuli are to the right of that decision boundary. Therefore, it is likely that some participants in the unidimensional category structure condition were also scoring less than 70%. If this were the case, I would have been able to select hypothetical participants from my simulation who were using a unidimensional or conjunction strategy but who were identified by the analysis as using a diagonal (GLC) strategy.

This is something that would be interesting to examine in future work using multidimensional scaling techniques. This would allow me to see how participants representations of stimuli change over time and how that would effect strategy identification.

From the perspective of COVIS, it is most interesting that it is possible to achieve high accuracy (above 92%) using a conjunction strategy that is misidentified as a diagonal (GLC) strategy. This means that it is possible that the participants in previous studies may

also have been using rule-based strategies that were misidentified as diagonal (GLC) strategies.

## 5.4 General discussion

The COVIS model of category learning is becoming influential and a great deal of evidence is cited in support of its predictions (Ashby & Maddox, 2011; Ashby & Valentin, 2016). Critically, the evidence for this model depends on correctly identifying the strategies that participants use to complete the learning tasks. COVIS hypothesises that its two systems can implement different types of strategy and so, each system can learn different types of category structure better. The importance of identifying strategies arises because the experiments investigating COVIS use this logic the other way around: they manipulate the category structures and hope that this encourages participants to use the optimum system, and thus strategy for that category structure. Of course, participants may continue to use the sub-optimum system for a particular category structure. Thus, identifying the strategies participants use is crucial: if the participants are using the correct strategy for that category structure, then the experimenters assume that they must also be using the correct learning system for that structure. Then, any differential effects of a manipulation on each category structure can be attributed to the existence of two systems of category learning, not differing numbers of sub-optimal responders.

Despite the importance of identifying participants' strategies, there is evidence to suggest that the analysis that researchers use within this literature is not up to the job. First, work by Donkin et al. (2015) showed that the output of the analysis varied depending on the category structure and the number of models included in the analysis. Second, the descriptions participants gave of their strategies in Chapters 2, 3 and 4 did not correspond well with the strategies identified by the model-based strategy analysis typically used in the COVIS literature. The key finding was that the majority of participants who learned an information-integration category structure reported using a rule-based strategy but were found by the model-based strategy analysis as using a diagonal (GLC) strategy, generally thought to signal implicit responding.

In this chapter, I conducted three model recovery simulations to demonstrate that the model-based strategy analysis used in the COVIS literature misidentifies the strategies participants use. In Section 5.2.1, I demonstrated that when participants are learning an information-integration category structure, whilst unidimensional and diagonal (GLC)

strategies are identified well, those using a conjunction strategy are as likely to be misidentified as using the optimum diagonal (GLC) strategy as identified as using the correct conjunction strategy. In Section 5.2.2, I demonstrated that when participants are learning a unidimensional category structure the strategy analysis will overestimate the number of optimum unidimensional responders; for this category structure, conjunction and diagonal (GLC) strategies are likely to be misidentified as a unidimensional strategy. Finally, I demonstrated that these systematic misidentifications are not just a theoretical issue, but could appear in practice. This would lead to alternative explanations of those results that are more consistent with single-process accounts. To do this, I simulated Experiment 2 of Smith et al. (2015) and showed that it was possible to reproduce their means, inferential statistics and strategy analysis using only participants who used rule-based strategies. This means that participants learned the information-integration category structure using a conjunction strategy but were found by the strategy analysis to be using a diagonal (GLC) strategy. This raises the possibility that participants in Experiment 2 of Smith et al. were also using rule-based strategies to learn the information-integration category structure.

### 5.4.1 Implications for the COVIS model

These simulations demonstrate the existence of an inferential flaw in the experiments argued to support COVIS: the strategy analysis is not accurate enough to act as a manipulation check. it cannot determine whether manipulating the category structure successfully elicited a corresponding switch in the categorisation system underlying participants' responses. Consequently, it becomes difficult to judge whether a particular COVIS-supporting dissociation is due to the existence of two distinct learning systems, or rather due to participants using different explicit strategies to learn each category structure. This means that the conclusions of a swathe of COVIS-supporting studies that relied on comparing rule-based and information-integration category structure have become uncertain.

More specifically, these simulations also have a knock on effect on a recent extension to this model-based approach called Iterative Decision Bound Modelling (iDBM Hélie, Turner, Crossley, Ell, & Ashby, 2016). iDBM aims to determine how participants' strategies change across learning by iteratively fitting decision bound models to participants trial by trial responses. Hélie et al. argue that this modelling procedure allows them to identify the strategies that participants are using on each trial, determine if participants

change strategy and estimate on what trial that occurs. They support these claims using several simulations and by applying their approach to data from Ell and Ashby (2006).

The simulations presented above raise several interesting points for this new work. First, their recovery rates for the simple strategy models (where there are no switches) appear to be much better than presented here (Hélie et al., 2016). This difference is likely due to the type of noise Hélie et al. added to their simulated responses. Hélie et al. added uniformly distributed noise: on any trial there was a certain probability of switching the category response from the one predicted by the strategy, to the other. In contrast, the simulations above added normally distributed perceptual and decisional noise as these types of noise are predicted by the strategy models. Therefore, the success of this type of model-based strategy analysis may depend on the type of noise actually experienced by participants. This is an experimental question that would be interesting to explore in future work.

Second, and possibly more importantly, they only included unidimensional, diagonal (GLC) and random models in their simulations and re-analysis of Ell and Ashby (2006). The simulations above (along with the experimental work reported in this thesis) demonstrate that these models alone are not sufficient to describe participants' behaviour in the category learning tasks used in the COVIS-supporting literature. Therefore, even if participants only added uniformly distributed noise and thus, decision bound modelling was more successful that estimated above, iDBM would need to be expanded in order to account for two-dimensional rule-based strategies such as a conjunction rule. Furthermore, the re-analysis of Ell and Ashby (2006) should be treated with caution as none of the data presented rules out the possibility that participants used complex, two-dimensional rule-based strategies to solve the tasks.

In relation to the experimental work in the previous chapters, the simulations I reported above also strengthen the evidence that participants were correctly reporting their categorisation strategies. In those experiments, participants learning information-integration category structures were consistently reporting using complex, rule-based strategies. In contrast, the model-based analysis identified these participants as using the correct (i.e. diagonal) strategy. In the above simulations, it was consistently shown that participants using a conjunction rule were likely to be misidentified as using a diagonal (GLC) strategy. Therefore, it seems plausible that all participants learn information-integration category

structures explicitly, using rule-based approaches.

The simulations above only partially speak to this point due to the simplicity of the strategy models considered. In the previous chapters, a substantial number of participants reported using explicit strategies that were more complex than a conjunction rule (the most complicated strategy typically modelled in the COVIS literature). It may be that the accuracy of strategy selection improves when these types of complex rule-based strategies are also included in the strategy analysis. Future work might explore a greater variety of rule-based models that better represent the strategies that are reported by participants.

### 5.4.2 Recommendations for future practice

This raises another feature of the strategy analysis that warrants further discussion: the role of the information criteria used to decide between strategy models. Specifically, I wish to discuss whether using information criteria (IC) that favour models with fewer parameters is sensible when determining the strategies that participants use. When IC are used to compare different theoretical models of category learning, there is no doubt they have a useful role (Wills & Pothos, 2012). In this case, if criteria were used that were neutral to the number of parameters involved, the winning model may win not because it best describes the underlying processes. Rather, it may win because it can better fit noise (also known as overfitting).

However, when considering strategy models for participants, perhaps the best fitting, simplest model might not best represent the strategies used by participants. Rather, the strategy model that fits the best *and* also corresponds well to the verbal reports given by participants might better represent the data. Consider a participant that used, and reported using, a conjunction strategy to score highly on an information-integration category structure. The simulations above indicate that it is likely that that participant would be identified as using a diagonal (GLC) strategy. In this type of case, it seems likely that the IC has over-emphasised reducing the number of parameters compared to the qualitative properties or the quantitative fit of the model. To avoid this possibility, it might be best to rely more on participants' verbal reports for the type of strategy and strategy models to quantitively specify that model.

Of course, this possibility is highly speculative and much work still needs to be done to test whether the verbal reports given by participants are also accurate. After all, it is possible that both the model-based strategy analysis and the verbal reports that participants

give do not represent the strategies that participants use. However, the outlook seems promising for verbal reports as previous work on the strategies participants use in the field implicit learning have found that verbal reports correspond well to other objective measures of learning (e.g., Lagnado et al., 2006; Lovibond & Shanks, 2002).

The possibility that the simplest model might not best represent the strategies used participants may have been exacerbated by the choice of the type of IC used in the COVIS literature. The most recent papers tend to use the Bayesian Information Criterion (e.g. Ashby & Vucovich, 2016; Spiering & Ashby, 2008) whilst earlier papers tend to use the Akaike Information Criterion (Ashby et al., 2002; Ell et al., 2009; Maddox, Bohil, & Ing, 2004; Maddox & Ing, 2005). Both these model selection criteria attempt to measure how well a strategy model balances fit with the number of parameters. Not only does this makes it difficult to compare the results of these analyses across experiments, but it also raises the possibility that the proportion of misidentified strategies may have increased. The BIC penalises the number of parameters in the strategy more than the AIC (Myung & Pitt, 1997). Therefore, it is possible that the switch towards using the BIC may be unwise: in this case, a more lenient criterion might be preferable (such as maximum likelihood).

Several other things may also improve the accuracy of the model-based analysis. Donkin et al.'s (2015) work suggests two improvements. First, including more complex models seems to improve the accuracy of the strategies identified by the strategy analysis. This recommendation is also supported by the numbers of complex two-dimensional rule-based strategies that were reported by participants in Chapters 2, 3 and 4. At the very least, the simulation above indicates that the strategy analysis should include conjunction strategies. Second, Donkin et al. also included a transfer phase that also appeared to improve the accuracy of the strategy analysis. For instance, using a transfer phase that included stimuli that filled the stimulus space would likely improve identification of strategies. However, the most important thing would be to conduct model-recovery simulations to demonstrate that for a particular experiment the strategy analysis can accurately identify the types of strategies reported by participants.

Of course, these thoughts are just hypothetical at present and require further investigation. However, if all these studies published their trial-by-trial level data, as is now recommended (Open Science Collaboration, 2015), these studies might be more future proof. This is especially an issue with the COVIS literature as it is not only the choice of IC

that varies across experiments. For example, the strategy analysis has been applied to different numbers of trials. For example, Ashby and Vucovich (2016) found the strategies for the last 100 trials for each participant, whereas in Spiering and Ashby (2008) they looked at blocks of 150 trials. Also, the models that are taken as indicative of the explicit or implicit category learning systems can vary between experiments. For example, the general quadratic classifier is argued to be an explicit rule-based strategy in Casale et al. (2012) but an implicit strategy in Ashby, Waldron, Lee, and Berkman (2001). If the data from these experiments were published, it would be possible to go back and conduct the same analysis for multiple papers. This would be highly desirable as ideas of best practice change across time.

### 5.4.3 Conclusions

In summary, the main take away messages are as follows. First, that we should be very careful about drawing firm conclusions from the results of this model-based analysis as currently put into practice. Second, the untrustworthiness of this strategy analysis casts doubt on the vast quantity of COVIS research that relies on it as a manipulation check. And finally, that there is more work to be done to determine whether there are any circumstances in which a decision boundary modelling technique may be reliable and valid.

# Chapter 6

# Conclusions

The work reported here examined the validity of the behavioural evidence for the dual-system model of category learning, COVIS. The COVIS model describes category learning using two parallel, competing, learning mechanisms (Ashby et al., 1998, 2011). The evidence for this model predominantly comes from experiments that compared learning of rule-based and information-integration category structures (Smith et al., 2015; Ashby & Valentin, 2016). This is a consequence of the assumption in this literature that the underlying learning system, participants' strategies and the category structure being learned are all linked. The Verbal System implements rule-based strategies and so optimally learns verbalisable rule-based category structures. The Procedural System can implement a variety of strategies and so is used to learn structures difficult to verbalise such as information-integration category structures. However, experiments in this literature use this logic in reverse. Researchers manipulate category structure to induce a corresponding switch in the learning system participants use. Then, they ascribe any differential effect of a secondary experimental manipulation on the learning of these category structures to the properties of the two systems supposedly underlying them.

The quantity of this type of evidence for the COVIS model appears to be large (Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2016). However, when re-examined by independent researchers, the support for the COVIS model is less clearcut (Newell et al., 2011). This is hardly surprising due to this literature's emphasis on using dissociation logic (Dunn & Kirsner, 2003; Newell et al., 2010). For the conclusions using this method to be valid, the rule-based and information-integration category structure conditions need to be matched on *everything* apart from their verbalisability. This is because it is the verbalisability of the category structure that COVIS hypothesises drives the change from the Verbal to the Procedural System (Ashby et al., 1998). Unfortunately for the COVIS model, several independent investigations have demonstrated that it is hard to control for extraneous variables using these category structures (Newell et al., 2011). The popularity of

the COVIS approach to category learning (Ashby & Valentin, 2016), along with the number of studies critiquing existing evidence (Newell et al., 2011) and the replication crisis phenomenon identified in psychology (Pashler & Wagenmakers, 2012), made it critical to investigate the remaining evidence for the COVIS model.

In the current work, I reported 6 experiments that critiqued two additional papers that have been argued to support the COVIS model. In Chapter 2, I re-examined a paper by Ashby et al. (2002) that looked at the effect of training type on the learning of unidimensional and information-integration category structures. Ashby et al. found that unidimensional category learning was unaffected by training type, whereas information-integration category learning was poorer with observational training compared to feedback training. They argued that this was because information-integration category learning is mediated by the Procedural System which requires reward prediction error to learn. However, I found that when I changed the rule-based category structure from a unidimensional structure to a conjunction one, participants in the rule-based condition also demonstrated significantly better learning with feedback training compared to observational training. Therefore, it seems the original dissociation was driven by the differing number of relevant dimensions for the discrimination, rather than the differing verbalisability of the category structures.

In Chapter 3, I re-examined a paper by Spiering and Ashby (2008) that found a differential effect of training order on rule-based and information-integration category learning. Counterintuitively, Spiering and Ashby found that participants learned an information-integration category structure better when they were initially trained on a harder version of the task compared to an easier version. For a rule-based category structure, they found no difference between training orders. Spiering and Ashby argued that this was because participants learning an information-integration structure who initially learned the easy stimuli could learn well using a simple rule and would therefore delay switching to the optimum Procedural System. However, participants that initially learned the hard stimuli would more quickly realise that the Verbal System was inadequate and switch to the optimum Procedural System. For the rule-based conjunction category structure, training order did not matter as the participants would begin learning by using the optimum system for the structure they were learning. In Chapter 3, I reported 4 experiments that all failed to replicate an advantage in information-integration category learning when initially training participants on the hard stimuli. If anything, participants learned the information-integration category structure better when being initially trained on the easy stimuli. This

demonstrates that we should be careful putting too much faith in a conclusion based on a single, un-replicated study.

These two strands of research found that two findings originally thought to support the COVIS model actually do not. They also cast slight doubt on one key assumption of the COVIS model: that the Procedural System produces "category knowledge opaque to declarative consciousness" (p. 2476 Smith et al., 2015). According to the logic of COVIS supporting experiments (although please see a more in depth discussion of this point below), participants who learn rule-based category structure should be able to verbalise how they did this, whereas those who learn an information-integration category structure should not. However, in Chapters 2 and 3, when I asked the participants to describe the strategy they used to sort the stimuli, nearly all of them could, regardless of the category structure they had learned. Furthermore, most of the participants who learned an information-integration category structure reported using complex verbalisable rules.

To directly investigate whether participants were in fact able to access category knowledge using declarative consciousness, in Chapter 4 I reported experiments that compared recognition memory performance between participants who had learned a rule-based or information-integration category structure. These experiments found no evidence that participants learning a rule-based category structure had more declarative memory than those learning an information-integration category structure. If anything, there was tentative evidence that participants in the information-integration category structure had superior levels of recognition memory performance. This is consistent with previous neuroscientific work that found superior activation in the medial temporal lobe for participants learning an information-integration category structure compared to a rule-based one (Carpenter et al., 2016). These experiments also reinforced the findings reported in Chapters 2 and 3. Contrary to the predictions of COVIS, this evidence suggests that participants use complex, somewhat idiosyncratic, rule-based strategies to learn information-integration category structures.

Proponents of COVIS might argue that my failure to find implicit responders was due to the fact that these participants failed to switch from the sub-optimum Verbal System to the optimum Procedural System. There are several objections to this criticism. First, if the same experimental paradigm worked in getting participants to switch in some situations, but not in others, this would likely occur in some of the published work supporting the

COVIS model too. Second, not only are these strategies not implicit, but a large quantity of them are also more complex than typically assumed to be implemented by the Verbal System of COVIS. This means that even if the participants presented here had merely failed to switch to the Procedural System, the Verbal System of COVIS also needs to be updated.

Most critically, though, these experiments met the criterion usually used in the COVIS literature to check that participants have indeed switched from the Verbal System to the Procedural. This criterion uses a model-based strategy analysis to check that participants are using the optimum strategy for the category structure they were assigned to learn. In other words, if the majority of participants in the information-integration category structure condition are found to be using a diagonal (GLC) strategy by this analysis, then they are assumed to be using the optimal Procedural System. This was indeed the case for all of my studies.

Comparing participants' verbal reports with the GRT strategy analysis raises an additional problem: they do not match. A large number of participants are identified as using the optimum diagonal (GLC) strategy but report using a complex, two-dimensional rule-based strategy. So which is right? In Chapter 5, I conducted several model-recovery simulations to test the validity of the model-based analysis as typically used in the COVIS literature. These simulations demonstrated that the strategy analysis is biased towards inappropriately confirming that the category structure manipulation has resulted in a corresponding shift in the strategies that participants used. When applied to responses to stimuli from an information-integration category structure, the model-based analysis overestimates the number of optimum diagonal (GLC) responses. As diagonal responders are assumed to be using the Procedural System, this misidentification results in researchers wrongly assuming that their category structure manipulation induced a corresponding change in the strategies participants use and therefore, the learning system in control of responding. Similarly, when the analysis is applied to responses to stimuli from a unidimensional category structure, the model-based analysis cannot discriminate between optimal unidimensional, conjunction and diagonal (GLC) strategies. Instead, participants are most likely to be identified as unidimensional responders no matter the strategy they actually used.

As a practical demonstration of the problem, I simulated Experiment 2 of Smith et al.

(2015). Smith et al. found an interaction between category structure type and response deadline conditions. Participants learning with a response deadline generally performed worse, however this difference was greater for participants learning the information-integration category structure than those learning the rule-based, unidimensional category structure. I demonstrated that it was possible to generate this pattern of results with participants that were exclusively using rule-based strategies. Additionally, I showed that the model-based analysis would misrepresent a quantity of these participants as using a diagonal (GLC) strategy, which matched the strategies identified by Smith et al.. Conceptually, this also makes sense: participants in the unidimensional condition were using simpler strategies than those in the information-integration condition. Smith et al.'s study merely indicates that time pressure has a greater effect on a more difficult task.

## 6.1 Implications for the COVIS model

The work presented in this thesis demonstrated that two papers in the literature argued to support COVIS do not hold up to further scrutiny. Chapter 2 showed that the experiments reported by Ashby et al. (2002) were better accommodated by a single-system approach. Whereas in Chapter 3 showed that an experiment by Spiering and Ashby (2008) failed to hold up to independent replication, the gold standard of psychological evidence (Ledgerwood, 2014; Pashler & Wagenmakers, 2012). They add to the growing literature that weakens the evidential support for COVIS (such as Newell et al., 2010; Nosofsky & Zaki, 2002; Nosofsky et al., 2005; Stanton & Nosofsky, 2007, 2013; Zaki & Kleinschmidt, 2014). However, these critiques are not comprehensive. Ashby and Valentin (2016) lists several additional studies that have not yet been re-examined. What does the evidence presented here say about them?

Some of the studies on Ashby and Valentin's (2016) list have indeed been independently examined, but have yet to be published. For example, Casale et al. (2012) examined whether participants who learned information-integration or unidimensional rule-based category structures could transfer that knowledge to another, similar categorisation task. They found that participants in the rule-based condition could, but participants in the information-integration condition had difficulties. However, Inkster, Edmunds, and Wills (In prep) noted that the nature of the transfer tasks differed between category structures and provided a simpler explanation based on selective attention. They found that it was possible to also demonstrate improvements in performance in a transfer task for both

rule-based and information-integration category structures. Other work has looked at a study that examined the effect of deferring feedback on rule-based and information-integration category learning. Smith et al. (2014) found participants could still learn rule-based category structures when feedback was deferred until the end of the block, but that performance on learning an information-integration "sharply" dropped. In contrast, Carpenter, Wills, Edmunds, and Milton (In prep) argued that this is due to Smith et al. using a unidimensional category structure rather than a conjunction, which would have the same number of relevant stimulus dimensions as the information-integration category structure (see Chapter 2). They found no evidence of a dissociation once the number of relevant stimulus dimensions was controlled for.

There are several more studies cited by Ashby and Valentin (2016) that may still hold evidence for the existence of two underlying mechanisms for category learning that, to my knowledge, have not been critiqued. Some of these findings are also simply explainable by single-system approaches and so provide limited additional evidence for the existence of two processes of category learning. For instance, Ashby et al. (1999) and Ell et al. (2012) examined whether participants could learn these rule-based and information-integration category structures in an unsupervised learning task. They found that participants were able to learn unidimensional rule-based category structures without feedback, but that they struggled to learn information-integration tasks that way. In contrast, previous evidence has found it possible for participants to learn versions of information-integration type category structures without feedback (e.g Love, 2002; Milton & Wills, 2004, 2009). Furthermore, these authors explained this behaviour using single-system approaches to category learning, meaning that these studies seem unlikely to hold the key to proving the existence of COVIS.

The remaining studies suffer from a more general problem: they rely too heavily on using rule-based and information-integration category structure tasks. For instance, Maddox et al. (2009) looked at the effect of sleep deprivation on information-integration category learning. Ell and Ashby (2006) compared the effect of category overlap on unidimensional rule-based and information-integration category learning, Maddox and colleagues looked at the effect of category discontinuity (Maddox, Filoteo, & Lauritzen, 2007; Maddox & Filoteo, 2011), Ell, Cosley, and McCoy (2011) examined the effect of stress, Nadler, Rabi, and Minda (2010) examined the effect of mood, Minda and Rabi (2015) examined the effect of ego depletion and so on. All these studies assume that rule-based category

structures tap into rule-based category learning and that information-integration category structures tap into implicit, procedurally-based category learning. Therefore, these authors conclude that the dissociations they found are due to the properties of these different systems.

In contrast, the evidence presented in this thesis indicates that this may not be the case. All the participants are able to report the strategies that they used, even for information-integration category structures. Furthermore, the types of strategies they report using appear to be rule-based. There is no reason to suppose that the participants in the COVIS-supporting papers are doing anything different. Especially, as none of these studies asked the participants which strategy they used. Therefore, the dissociations reported in these experiments are most likely due to the secondary manipulation (such as mood, stress etc) interacting with the strategies the participants use. Of course, this speculation needs to be experimentally examined. Therefore, crucial for the COVIS model is future work that re-examines these COVIS supporting studies.

## 6.2 Neuropsychological evidence for COVIS

Some may argue that the largest source of support for the COVIS model comes from neuropsychological work (Ashby et al., 1998, 2011). For example, studies have looked at participants diagnosed with Parkinson's disease (for reviews see Filoteo & Maddox, 2007; Price, Filoteo, & Maddox, 2009) or anorexia, as well as some patients with brain lesions. Neuropsychological dissociations have been argued to be critical evidence for the existence of dual-systems approaches to category learning and other areas of cognition such as memory (Squire, 1992, although see Berry et al., 2012; Kinder & Shanks, 2003). This leads to the question: does any of the evidence presented in this thesis lead us to re-evaluate the neuropsychological support for COVIS?

### 6.2.1 Parkinson's disease

A large quantity of work has examined category learning of participants with Parkinson's disease (for reviews see Filoteo & Maddox, 2007; Price et al., 2009). Patients with Parkinson's disease are particularly interesting to look at in relationship to the COVIS model due to the role dopamine plays in the model (Ashby & Valentin, 2016). Switching between rules in the Verbal System is argued to be closely related to dopamine levels in the basal ganglia (Ashby et al., 2011). As Parkinson's patients have abnormally low levels of dopamine in the striatum, these patients allow testing of that assumption (Ashby

& Valentin, 2016). Indeed, much evidence has been argued to support the assumptions of the Verbal System of COVIS (Filoteo, Maddox, & Others, 2007; Ell, Weinstein, & Ivry, 2010; Price et al., 2009). However, these studies only speak to the properties of a rule-based system of category learning and do not necessarily require the existence of a second implicit, Procedural System.

That being said, several studies have examined implicit category learning (as indexed by information-integration category learning) with Parkinson's patients (Filoteo & Maddox, 2007). However, the results of these studies have been mixed. Maddox and Filoteo (2001) demonstrated that Parkinson's patients had deficits learning a non-linear information-integration task, finding no difference for a linear information-integration task. In contrast, (Ashby, Noble, Filoteo, Waldron, & Ell, 2003) found a dissociation in performance between learning rule-based and information-integration tasks. Patients with Parkinson's performed more poorly than age-matched controls in the unidimensional rule-based task, but the same as controls in the (linear) information-integration category structure task. However, all the patients in these studies were taking some form of dopaminergic medication which has argued to affect the accuracy of conclusions about deficits in learning in patient with Parkinson's disease (Newell et al., 2011). Therefore, it appears that the evidence for a dual-systems approach to category learning from patients with Parkinson's disease is limited.

### 6.2.2 Anorexia nervosa

The COVIS model has also been argued to predict patterns of responding in patients diagnosed with anorexia nervosa. Filoteo et al. (2014) examined set shifting behaviour in weight-restored anorexia nervosa patients. They found that the Verbal System predicted the inability of these patients to switch rules. However, Filoteo et al. they only examined performance on rule-based category learning tasks. Therefore, this study does not require a secondary, implicit Procedural System.

### 6.2.3 Lesions

Schnyer et al. (2009) examined rule-based and information-integration category learning in patients with ventral pre-frontal cortex lesions. They found that participants who were identified by the model-based analysis as using sub-optimum strategies were also found to be impaired at other tasks such as the Wisconsin Card Sorting Test. Donkin et al. (2015) re-examined their analysis and found that both the category structure and

the number of strategy models included in the analysis biased the results of this experiment. Furthermore, Schnyer et al. did not find a dissociation between rule-based and information-integration category structures and so finds very little evidence for a dual-system approach to category learning.

All-in-all, the neuropsychological evidence appears to suffer from the same limitations as the behavioural work with non-patient populations. As a general rule, they rely too heavily on rule-based and information-integration category structures without checking the strategies that participants actually used to complete the task.

### 6.2.4 fMRI

Some research has also attempted to look at the brain regions associated with learning different category structures using fMRI. For instance, Nomura et al. (2007) found that participants who learned the rule-based category structure had more activity in the medial temporal lobe and less activity in the caudate body than those who learned an information-integration structure. Nomura and Reber (2008) reanalysed this data taking into account the participants' strategies identified by the model-based analysis. They found more activation in the prefrontal cortex and less activation in the occipital cortex for those identified as using a rule-based strategy compared to those identified as using a diagonal strategy. However, recent work by Carpenter et al. (2016) showed that when the category structures were better matched for number of relevant dimensions, error rates and category separation, there was considerable overlap in brain activation between the rule-based and information-integration category structures. Also, contrary to the predictions of COVIS, Carpenter et al. found greater activation in the medial temporal lobe in the information-integration condition compared to the rule-based condition. This indicates that the initial dissociations in brain regions were due to extraneous variation between the category structures, not the existence of two systems of category learning.

## 6.3 Verbal reports

The work presented in this thesis (and the COVIS-supporting studies briefly reviewed above) has highlighted the importance of determining the strategies that participants used to learn a particular category structure. Here, I have demonstrated that a simple strategy questionnaire can provide a rich source of data and an easy manipulation check for the model-based approach typically used in the COVIS literature. The verbal reports also indicated that the range of strategies participants report is much broader than commonly

supposed in the COVIS literature. However, some may argue that questionnaires may not be the best way of determining participants' response strategies (Newell & Shanks, 2014).

The first such objection is theoretically motivated. The COVIS model predicts that participants will learn easy-to-verbalise category structures using the explicit Verbal System, and hard-to-verbalise category structures using the implicit Procedural System (Ashby et al., 1998, 2011). This is because the Verbal System mediates easy-to-verbalise strategies and the Procedural System mediates all the other types of strategy. Clearly, COVIS would predict that when participants learn a rule-based category structure they would be able to describe the strategy they were using. This was indeed the case for all the rule-based category structures reported in the previous chapters.

Under this account, where it becomes difficult to interpret verbal reports is for those learning information-integration category structures. Here, COVIS might predict one of three types of verbal report. One possibility is that participants might decline to report a verbalisable strategy, instead reporting that they guessed or did not know. This is typically taken as a measure of implicit responding in other work exploring implicit learning (e.g., Yeates, Wills, et al., 2012; Yeates et al., 2013; Vadillo, Konstantinidis, & Shanks, 2015). My data speak against the possibility that participants were learning information-integration category structures implicitly and failing to report strategies, as equal numbers of participants failed to report strategies in both types of category structure condition.

Another possibility is that participants might be predicted to report responding on the basis of a gut feeling when they are using the Procedural System. This hypothesis would be consistent with with the account of dual-system decision making as popularised by the books "Thinking, fast and slow" by Kahneman (2011) or "Blink" by Gladwell (2005). My data also speaks against this as only one participant reported something that might be described as "going with their gut" and that participant was in a rule-based condition (see Experiment 7).

A final possibility is that participants implicitly learning an information-integration category structure incorrectly report a rule-based strategy. Perhaps, this might happen because they cannot express the strategy they are using but do not want to say nothing due to the experimental demands. This is a more difficult possibility to rule out. However, the types of strategy descriptions given by participants in both rule-based and information-

integration category structure tasks in my experiments are very similar. Also, for the experiments reported here, the verbal reports given by participants who learned a rule-based category structure correspond well with the strategies found by the model-based analysis. It seems unlikely (although not impossible) that the verbal reports are accurate for one category structure and not another. That, and the fact that no evidence of dual-systems of category learning was found, suggests that a reasonable interim assumption is that participants in all conditions are using verbalisable rules.

A second objection to relying on verbal reports might be that verbal reports are inherently unreliable because they are subjective. For instance, in the abstract of Nisbett and Wilson (1977), they stated that participants have "little or no introspective access to higher order cognitive processes" (p.231, Nisbett & Wilson, 1977). However, the actual state of affairs may be more nuanced, with the accuracy of verbal report measures critically depending on what is being reported. Indeed, Nisbett and Wilson acknowledge that experiments studying learning are well designed to elicit accurate verbal reports. This claim has been born out in the literature. For instance, Lovibond and Shanks (2002) found that participants' verbal reports of what they learned in the task matched exceedingly well to the behavioural measures. Similarly, Lagnado et al. (2006) found that participants in a multiple-cue learning paradigm could accurately report the task structure and their own judgement processes. These successful uses of verbal reports in previous learning studies provides support that the verbal reports discussed here may also be valid.

That being said, in the studies by Lovibond and Shanks (2002) and Lagnado et al. (2006), it was possible to make strong claims as to the accuracy of the verbal reports because they were validated by independent behavioural or computational data (Locke, 2009). Unfortunately, this was not possible in the current work, as the leading cognitive modelling strategy was shown to not be reliable, especially for the information-integration category structure (see Chapter 5). That being said, providing a more reliable, objective measure of strategies for the category structures used in the COVIS literature is of paramount importance. As is objectively verifying the validity of participants' verbal reports. Perhaps by using eye-tracking, fMRI or more subtle computational techniques such as multidimensional scaling, future work would be able to better quantify the strategies that participants are using.

As well as these more serious arguments, it is important to note that verbal reports can

be somewhat vague. For example, if a participant reports using a unidimensional strategy based on the size of the stimuli, it is difficult to determine from that report alone which stimuli count as “big” and which “small.” It also appears that these strategies are descriptions of the participants’ patterns of responding, rather than the underlying representational system mediating learning. In other words, the verbal reports may describe well how a particular person represented the task to others, but these descriptions are only of limited use in describing the underlying mechanism of category learning; one description would be consistent with several formal models of category learning. For example, a participant who described using a unidimensional strategy might be best described in one of several ways. For the unidimensional category structures used in the COVIS literature, this verbal report would be consistent with the pattern of responding from an exemplar model (e.g. Kruschke, 1992), a prototype model (e.g. Rosch & Mervis, 1975), a model that fits a linear boundary (e.g. Ashby & Gott, 1988) or one that generates clusters as necessary (e.g. Love, Medin, & Gureckis, 2004). Therefore, future work could investigate how well verbal reports correspond to these different types of models by comparing multiple models of category learning to a single data set, including the reports given by participants (as advocated in Wills & Pothos, 2012; Wills, O’Connell, Edmunds, & Inkster, Accepted).

## 6.4 Limitations

One limitation is that, although these experiments reported above have demonstrated that a dual-system model is not required, very little attention has been given to what type of category learning model would best describe the pattern of results demonstrated here. Much of the evidence would be consistent with a broadly rule-based approach (such as suggested by Milton, Wills, & Hodgson, 2009; Wills et al., 2013). In this approach, both rule-based and information-integration category structures are learned by participants using rule-based strategies. The strategies that participants use to learn information-integration category structures are generally more complex than those used to learn rule-based category structures. Then, any emerging dissociations are due to the fact that simple rules are easier to implement than more complicated ones (Milton & Wills, 2004; Wills et al., 2015).

Furthermore, the experimental phenomena reported in this thesis lack proper theoretical explanation. For example, in Chapter 2 I demonstrated that the effect of training type

was not dissociable between rule-based and information-integration once the number of relevant dimensions was controlled for. However, I failed to explain sufficiently why feedback training was superior to observational training for both these category structures. Therefore, future work should aim to investigate these phenomena outside of the COVIS framework.

## 6.5 Conclusion

In summary, I would like to advocate a move towards again assuming that category learning is mediated by a single system. Furthermore, that we should closely examine any evidence that supports the existence of multiple systems of category learning to ensure that we are not making false inferences based on the limitation of our experimental methods.

# Appendix A

# Model-recovery procedure

The trial-level raw data and code will be available at `www.willslab.co.uk/ply34` once this work is published.

Participants' responses can be subject to several types of sub-optimality (Maddox & Ashby, 1993). The one that is most crucial for the experimental COVIS literature is choosing the wrong strategy type. In this analysis, I looked to see whether the correct strategy type was selected from responses generated from three common models: unidimensional, conjunction and diagonal (GLC). These three strategy models are the ones most often included in the model-based strategy analysis used in the COVIS literature. The unidimensional strategy assumes that participants make categorisation judgements based on a single dimension rule. The conjunction strategy assumes that participants make categorisation judgements based on two dimensions and it corresponds to a logical conjunction rule. The diagonal strategy is the optimum strategy for the information-integration category structure and assumes that participants make the two stimulus dimensions commensurate and combine them to make a judgement.

Second, participants may find it difficult to identify the perceptual features of the stimuli, i.e. their responses may be subject to perceptual noise. To simulate this kind of sub-optimality, perceptual noise was applied to each stimulus as a bivariate, normal distribution centred on the stimulus. The variance on each stimulus dimension was equal and there was no correlation between the dimensions.

Third, participants may apply their chosen strategy sub-optimally. In other words, there may be variance in where they apply the decision bound. Decisional noise corresponds to variation in where the decision boundary passing through stimulus space is supposed to lie. This was modelled as unidimensional normal distribution, orthogonal to the decision boundary. Both perceptual and decisional noise are central tenets of General Recognition Theory (Ashby & Gott, 1988; Ashby & Soto, 2015).

## A.1 Category structures

These simulations used the category structures from Smith et al. (2015). Here, the coordinates for each stimulus are generated by sampling from two bivariate normal distributions, one for each category. This method of creating category structures was proposed by Ashby and Gott (1988). These category structures were selected as they are representative of the most common types of structure in the COVIS literature (Ashby & Valentin, 2016) and so seem a fair test of whether using strategy analysis as a manipulation check works.

I sampled 200 points from each normal distribution, defined according to the parameters in Table A.1.

*Table A.1:* Parameters to define the category structures used these simulations. Originally used in Smith et al. (2015).

| Task and category | $\text{Mean}_X$ | $\text{Mean}_Y$ | $\text{Var}_X$ | $\text{Var}_Y$ | $\text{Cov}_{XY}$ |
|---|---|---|---|---|---|
| Unidimensional | | | | | |
| Category A | 35.86 | 50.0 | 16.33 | 355.55 | 0 |
| Category B | 64.4 | 50.0 | 16.33 | 355.55 | 0 |
| Information-integration | | | | | |
| Category A | 40.0 | 60.0 | 185.94 | 185.94 | 169.61 |
| Category B | 60.0 | 40.0 | 185.94 | 185.94 | 169.61 |

## A.2 Response generating models

There were three generating strategies in these simulations: unidimensional, conjunction and the diagonal, general linear classifier.

For the *unidimensional strategy* the responses are simulated by a linear category boundary perpendicular to the $x$-axis. Then, stimuli that fell to the left of the boundary were assigned 'Category A' responses and those that fell on the other side were given 'Category B' responses. In all the simulations, the boundary was assumed to be at $x = 50.0$.

For the *general linear classifier strategy* the responses were simulated by a linear category boundary diagonally dividing the stimulus space. Then, the stimuli that fell above the boundary were given 'Category A' responses and those below 'Category B' responses. In these simulations, the diagonal boundary was fit to the category structure, so as to score the highest accuracy.

For the *conjunction strategy* the responses were simulated by two linear boundaries, one

parallel and one perpendicular to the $x$-axis, that separated off a section of the stimulus space. Then, the stimuli that fell in that 'corner' of the space were given 'Category A' responses and the others 'Category B' responses. The position of the boundaries were fit to the category structure, so as to score the highest accuracy.

Two different types of noise were added, at various levels, to the responses given by this model. The first type was *perceptual noise*. This was modelled as a normal distribution with a mean centred on the stimuli and standard deviation $\sigma$. This took the values 0, 5, 10, 15 or 20. For the unidimensional and general linear classifier strategies this had the effect of varying the stimulus coordinates perpendicular to the decision boundary. For the conjunction strategy, the stimuli varied in both the $x$- and $y-$ directions with a covariance of 0.

The second type of noise was *decisional noise*. This was also modelled as normal distribution with the mean as the category boundary and standard deviation $\sigma$ representing a shift of the whole boundary in stimulus space. The standard deviation could take values of 0, 12.5, 25, 37.5 and 50. The final response for each stimulus was determined by seeing which side of the boundary (or boundaries in the case of the conjunction strategy) with decisional noise, the stimulus value with added perceptual noise was. For each strategy and category structure, at each level of noise, 20 participants were generated, 3000 in total.

## A.3 Model-based strategy analysis

Once the simulated responses had been generated, four types of models were fitted using the model-based strategy analysis used in the COVIS literature (Maddox & Ashby, 1993). The models were unidimensional, general linear classifier, conjunction and random.

The *unidimensional strategy* model was assumed to have two parameters: the value of the boundary and perceptual noise. Additionally, though only a unidimensional rule based on the $x$-axis was used to generate responses, both types of unidimensional rule were included in the model-fitting procedure. This was to account for the possibility in the information-integration category structure that either the rule based on $x$- or $y$-axis might best fit the data.

The *general linear classifier strategy* model was assumed to have three parameters: the gradient and intercept of the boundary as well as perceptual noise.

The *conjunction strategy* model was assumed to have four parameters: the values of the two decision boundaries and two noise parameters, one for each stimulus dimension.

The *random strategy* model assumed that for each stimuli category membership was assigned at random and has a single parameter: the probability of any stimulus being assigned to Category A.

As in Smith et al. (2015) the strategy that best represents each hypothetic participant's responding was defined as the one that minimises the Bayesian Information Criterion (BIC; Schwarz, 1978). This is calculated as follows:

$$BIC = -2\ln L + k\ln N \tag{A.1}$$

where $L$ is the maximum likelihood of the model, $k$ is the number of parameters and $N$ is the number of data points. This criterion represents a trade off between how well the model fits and the number of parameters (the parsimony of the model).

Additionally, we calculated the Schwarz weight for each winning model (Wagenmakers & Farrell, 2004):

$$w_i(BIC) = \frac{\exp\{-\frac{1}{2}\Delta_i(BIC)\}}{\sum_{k=1}^{K}\exp\{-\frac{1}{2}\Delta_K(BIC)\}} \tag{A.2}$$

where $\Delta_i(BIC)$ is the difference between the BIC of model $i$ and the smallest BIC value of all the models fitted. The denominator is summed over $K$ models. Schwarz weights can be interpreted as the probability that this model is the best model for that participant.

## References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi: 10.1109/TAC.1974.1100705

Ashby, F. G. (2014). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychonomic Bulletin & Review*, *21*(4), 935–946. doi: 10.3758/s13423-013-0578-x

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., & Crossley, M. J. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews: . . .*, *3*, 363–376. doi: 10.1002/wcs.1172

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*, 1114–1125. doi: 10.3758/BF03196132

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53. doi: 10.1037/0278-7393.14.1.33

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., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677. doi: 10.3758/BF03196423

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

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89. doi: 10.1016/j.tics.2004.12.003

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). New York, NY: Cambridge University Press.

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

doi: 10.3758/BF03207622

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. In *Elementary cognitive mechanisms* (pp. 13–34).

Ashby, F. G., & Soto, F. A. (2016). The neural basis of General Recognition Theory. In J. W. Houpt & L. M. Blaha (Eds.), *Mathematical models of perception and cognition volume ii: A festschrift for james t. townsend* (pp. 1–28). Routledge.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179. doi: 10.1037/0033-295X.93.2.154

Ashby, F. G., & Valentin, V. V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 547–572). Elsevier.

Ashby, F. G., & Valentin, V. V. (2016). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 547–572). New York, NY: Elsevier.

Ashby, F. G., & Vucovich, L. E. (2016). The role of feedback contingency in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0000277

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363–378. doi: 10.3758/BF03210826

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77–96. doi: 10.1037/0096-3445.130.1.77

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175.

Baguley, T., & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology*, *63*, 695–698.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181. doi: 10.1016/0022-2496(79)90016-6

Bayes, & Price. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, *53*, 370–418. doi: 10.1098/

rstl.1763.0053Berger,

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. A. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review*, *119*, 40–79. doi: 10.1037/a0025464

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*. doi: 10.1002/hbm.23259

Carpenter, K. L., Wills, A. J., Edmunds, C. E. R., & Milton, F. (In prep). Does deferred feedback sharply dissociate implicit and explicit category learning?

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449. doi: 10.3758/s13421-011-0154-4

Church, B. A., Mercado, E., Wisniewski, M. G., & Liu, E. H. (2013). Temporal Dynamics in auditory perceptual learning: Impact of sequencing and incidental Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 270–276. doi: 10.1037/a0028647

Cohen, J. (1998). *Statistical power analysis for the behavioural sciences*. Lawrence Erlbaum Associates.

Conroy, M. A., Hopkins, R. O., & Squire, L. R. (2005). On the contribution of perceptual fluency and priming to recognition memory. *Cognitive, Affective & Behavioral Neuroscience*, *5*, 14–20. doi: 10.3758/CABN.5.1.14

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. doi: 10.1037/0096-3445.104.3.268

Crossley, M. J., Paul, E. J., Roeder, J. L., & Ashby, F. G. (2015). Declarative strategies persist under increased cognitive load. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015-0867-7

Davis, T., Love, B. C., & Preston, A. R. (2012). *Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI.* (Vol. 38; Tech. Rep. No. 4). Retrieved from `http://tylerdavisphd.com/papers/TDavis.pdfpapers3://publication/uuid/B57D5A18-1E4D-4C80-8339-2CA475A2A012` doi:

10.1037/a0027865

DeCaro, M. S., Carlson, K. D., Thomas, R. D., & Beilock, S. L. (2009). When and how less is more: reply to Tharp and Pickering. *Cognition*, *111*(3), 397–403. doi: 10.1016/j.cognition.2009.03.002

DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: sometimes less working memory capacity is better than more. *Cognition*, *107*, 284–94. doi: 10.1016/j.cognition.2007.07.001

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290. doi: 10.1177/1745691611406920

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., Nosofsky, R. M., Donkin, C., . . . Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 933–948. doi: 10.1037/xlm0000083

Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby (2014). *Psychonomic Bulletin & Review*, *21*, 947–54. doi: 10.3758/s13423-014-0637-y

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex*, *39*, 1–7. doi: 10.1016/S0010-9452(08)70070-4

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 840–859. doi: 10.1037/a0027867

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*, 1203–1222. doi: 10.1080/17470218.2014.978875

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*, 1013–1026. doi: 10.3758/BF03193362

Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *The Quarterly Journal of Experimental Psychology*, *65*, 1537–1562. doi: 10.1080/17470218.2012.658821

Ell, S. W., Cosley, B., & McCoy, S. K. (2011). When bad stress goes good: increased threat reactivity predicts improved category learning performance. *Psychonomic Bulletin & Review*, *18*(1), 96–102. doi: 10.3758/s13423-010-0018-0

Ell, S. W., Ing, A. D., & Maddox, W. T. (2009). Criterial noise effects on rule-based category learning: The impact of delayed feedback. *Attention, Perception, & Psychophysics*, *71*, 481–489. doi: 10.3758/APP.71.6.1263

Ell, S. W., Weinstein, A., & Ivry, R. B. (2010). Rule-based categorization deficits in focal basal ganglia lesion and Parkinson's disease patients. *Neuropsychologia*, *48*, 2974–2986. doi: 10.1016/j.neuropsychologia.2010.06.006

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. doi: 10.1207/s15516709cog1402_1

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140. doi: 10.1037/h0045156

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*, 75–89. doi: 10.1016/S0022-2496(02)00025-1

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*, 415–423. doi: 10.1177/0956797610362646

Filoteo, J. V., & Maddox, W. T. (2007). Category Learning in Parkinson's Disease. In M.-K. Sun (Ed.), *Research progress in alzheimer's disease and dementia* (Vol. 3, pp. 339–365). Hauppauge, NY: Nova Science Publishers, Inc.

Filoteo, J. V., Maddox, W. T., & Others. (2007). Characterizing rule-based category learning deficits in patients with Parkinson's disease. *Neuropsychologia*, *45*(2), 305–320. doi: 10.1016/j.neuropsychologia.2006.06.034

Filoteo, J. V., Paul, E. J., Ashby, F. G., Frank, G. K. W., Hélie, S., Rockwell, R., ... Kaye, W. H. (2014). Simulating category learning and set shifting deficits in patients weight-restored from anorexia nervosa. *Neuropsychology*, *28*, 741–751. doi: 10.1037/neu0000055

Gabrieli, J. D. E., & Fleischman, D. (1995). Double dissociation between memory systems underlying explicit and implicit memory in the human brain. *Psychological Science*, *6*, 76–82. doi: 10.1111/j.1467-9280.1995.tb00310.x

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experi-*

*mental Psychology*, *50*(1), 114–122. doi: 10.1037/1196-1961.50.1.114

Gladwell, M. (2005). *Blink: The power of thinking without thinking*. New York: Little, Brown and Co.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. doi: 10.3758/s13421-011-0174-0

Haslam, C., Hodder, K. I., & Yates, P. J. (2011). Errorless learning and spaced retrieval: How do these methods fare in healthy and clinical populations? *Journal of Clinical and Experimental Neuropsychology*, *33*(4), 432–447. doi: 10.1080/13803395.2010.533155

Heathcote, A., Brown, S. D., & Wagenmakers, E.-j. (2014). *An Introduction to Good Practices in Cognitive Modeling*. New York, NY: Springer.

Hélie, S., & Ashby, F. G. (2012). Learning and transfer of category knowledge in an indirect categorization task. *Psychological Research*, *76*(3), 292–303. doi: 10.1007/s00426-011-0348-1

Hélie, S., Turner, B. O., Crossley, M. J., Ell, S. W., & Ashby, F. G. (2016). Trial-by-trial identification of categorization strategy using iterative decision-bound modeling. *Behavior Research Methods*. doi: 10.3758/s13428-016-0774-5

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*. doi: 10.1037/a0037989

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*, 72–76. doi: 10.1016/j.jarmac.2014.04.009

Inkster, A. B., Edmunds, C. E. R., & Wills, A. J. (In prep). Analogical transfer in rule-based and information-integration categorization.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852–857. doi: 10.3758/BF03196776

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kinder, A., & Shanks, D. R. (2003). Neuropsychological dissociations between priming

and recognition: a single-system connectionist account. *Psychological Review*, *110*, 728–44. doi: 10.1037/0033-295X.110.4.728

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. doi: 10.1037/a0015729

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi: 10.1037/0033-295X.99.1.22

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603. doi: 10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS and Stan* (2nd ed.). Academic Press.

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

LaShell, P. J. (2010). *What's in a rule: Two-dimensional rule use in category learning* (Doctoral dissertation, UC Riverside: Psychology). Retrieved from `https://escholarship.org/uc/item/70x2x6wk`

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology*, *45*, 511–516. doi: 10.1037/h0057135

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*(3), 275–277. doi: 10.1177/1745691614529448

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*, 266–282. doi: 10.3758/s13421-014-0458-2

Lewandowsky, S., Palmeri, T. J., & Waldmann, M. R. (2012). Introduction to the Special Section on theory and data in categorization: Integrating computational, behavioral, and cognitive neuroscience approaches. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 803–806. doi: 10.1037/a0028943

Lewandowsky, S., Yang, L.-X., Newell, B. R., & Kalish, M. L. (2012). Working memory does not dissociate between different perceptual categorization tasks.

*Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 881–904. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027298papers3://publication/doi/10.1037/a0027298` doi: 10.1037/a0027298

Locke, E. A. (2009). It's Time We Brought Introspection Out of the Closet. *Perspectives on Psychological Science*, *4*, 24–25. doi: 10.1111/j.1745-6924.2009.01090.x

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*, 835–863. doi: 10.1037/0033-295X.111.4.835

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*, 829–835. doi: 10.3758/BF03196342

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*, 309–332. doi: 10.1037/0033-295X.111.2.309

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of experimental psychology. Animal behavior processes*, *28*, 3–26. doi: 10.1037/0097-7403.28.1.3

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. New Jersey, US: Lawrence Erlbaum Associates.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*, 354–74. doi: 10.3758/BF03206893

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70. doi: 10.3758/BF03211715

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. doi: 10.3758/BF03195849

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-

based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*, 945–952. doi: 10.3758/BF03196726

Maddox, W. T., & Filoteo, J. V. (2001). Striatal contributions to category learning: Quantitative modeling of simple linear and complex nonlinear rule learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society*, *7*, 710–727. doi: 10.1017/S1355617701766076

Maddox, W. T., & Filoteo, J. V. (2011). Stimulus range and discontinuity effects on information-integration category learning and generalization. *Attention, Perception, & Psychophysics*, *73*, 1278–1295. doi: 10.1017/CBO9781107415324.004

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 227–245. doi: 10.1037/0278-7393.30.1.227

Maddox, W. T., Filoteo, J. V., & Lauritzen, J. S. (2007). Within-category discontinuity interacts with verbal rule complexity in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 197–218. doi: 10.1037/0278-7393.33.1.197

Maddox, W. T., Glass, B. D., Wolosin, S. M., Savarie, Z. R., Bowen, C., Matthews, M. D., & Schnyer, D. M. (2009). The effects of sleep deprivation on information-integration categorization performance. *Sleep*, *32*, 1439–1448.

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., Lauritzen, J. S., & Ing, A. D. (2007). Cognitive complexity effects in perceptual classification are dissociable. *Memory & Cognition*, *35*(5), 885–894. doi: 10.3758/BF03193463

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback in perceptual category learning. *Cognition*, *108*, 578–589. doi: 10.1016/j.cognition.2008.03.010.

Magnussen, S. (2000). Low-level memory processes in vision. *Trends in Neurosciences*, *23*(6), 247–251. doi: 10.1016/S0166-2236(00)01569-1

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research:

How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. doi: 10.1177/1745691612460688

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*.

Matsuki, K. (2014). grt: General recognition theory (0.2) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=grt`

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. doi: 10.1037/0033-295X.85.3.207

Meeter, M., Myers, C. E., Shohamy, D., Hopkins, R. O., & Gluck, M. A. (2006). Strategies in probabilistic categorization: Results from a new way of analyzing performance. *Learning & Memory*, *13*, 230–239. doi: 10.1101/lm.43006

Milton, F., Longmore, C. a., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 676–692. doi: 10.1037/0096-1523.34.3.676

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*, 1326–1336. doi: 10.1111/j.1460-9568.2011.07847.x

Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 407–415. doi: 10.1037/0278-7393.30.2.407

Milton, F., & Wills, A. J. (2009). Long-term persistence of sort strategy in free classification. *Acta Psychologica*, *130*, 161–167. doi: 10.1016/j.actpsy.2008.12.001

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *NeuroImage*, *46*, 319–326. doi: 10.1016/j.neuroimage.2009.01.043

Minda, J. P., & Rabi, R. (2015). Ego depletion interferes with rule-defined category learning but not non-rule-defined category learning. *Frontiers in Psychology*, *6*, 1–9. doi: 10.3389/fpsyg.2015.00035

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95. doi: 10.3758/BF03210778

Nadler, R. T., Rabi, R., & Minda, J. P. (2010). Better mood and better performance. Learning rule-described categories is enhanced by positive mood. *Psychological*

*Science*, *21*(12), 1770–1776. doi: 10.1177/0956797610387441

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.38.5.563

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In *Psychology of learning and motivation* (Vol. 54, pp. 167–215). doi: 10.1016/B978-0-12-385527-5.00006-1

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. doi: 10.3758/BF03194098

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves implicit perceptual categorization: A comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*, 386–9. doi: 10.1177/0956797612457387

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*, 1–19. doi: 10.1017/S0140525X12003214

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. doi: 10.1037/1089-2680.2.2.175

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi: 10.1037/0033-295X.84.3.231

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., ... 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. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. doi:

10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. doi: 10.1037/0278-7393.14.4.700

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79. doi: 10.1037/0033-295X.101.1.53

Nosofsky, R. M., Stanton, R. D., & Zaki, S. R. (2005). Procedural interference in perceptual classification: implicit learning or cognitive complexity? *Memory & Cognition*, *33*, 1256–1271. doi: 10.3758/BF03193227

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*(4), 247–255. doi: 10.1111/1467-9280.00051

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940. doi: 10.1037/0278-7393.28.5.924

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, *132*, 491–511. doi: 10.1037/0096-3445.132.4.491

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). Retrieved from `http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/26315443` doi: 10.1126/science.aac4716

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 548–568. doi: 10.1037/0278-7393.21.3.548

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi: 10.1177/1745691612465253

Paul, E. J., & Ashby, F. G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*,

*7*(177), 1–17. doi: 10.3389/fncom.2013.00177

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Plummer, M. (2003). *Jags: A program for analysis of bayesian graphical models using gibbs sampling.*

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–67. doi: 10.1037/a0033194

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

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raijmakers, M. E. J., Dolan, C. V., & Molenaar, P. C. M. (2001). Finite mixture distribution models of simple discrimination learning. *Memory & Cognition*, *29*, 659–677. doi: 10.3758/BF03200469

Raijmakers, M. E. J., Jansen, B. R. J., & van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review*, *24*, 289–321. doi: 10.1016/j.dr.2004.06.002

Roeder, J. L., & Ashby, F. G. (2016). What is automatized during perceptual categorization? *Cognition*, *154*, 22–33. doi: 10.1016/j.cognition.2016.04.005

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. doi: 10.1016/0010-0285(75)90024-9

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*, 534–553. doi: 10.1037/0096-3445.133.4.534

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia*, *47*, 2995–3006. doi: 10.1016/j.neuropsychologia.2009.07.011

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi: 10.1214/aos/1176344136

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18*, 414–421. doi: 10.3758/s13423-010-0047-8

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. a., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447–57. doi: 10.1177/0956797613509112

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*, 2476–2490. doi: 10.3758/s13414-015-0933-2

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, *19*, 1169–1177.

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, *4*, 232–243. doi: 10.1017/CBO9781107415324.004

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*, 171–177. doi: 10.1016/j.nlm.2004.06.005

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, *8*, 872–883. doi: 10.1038/nrn2154

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. doi: 10.3758/BF03193507

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1174–1191. doi: 10.1037/a0031670

Suret, M., & McLaren, I. P. L. (2003). Representation and discrimination on an artificial dimension. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *56*, 30–42. doi: 10.1080/02724990244000142

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*(3), 411–415. doi: 10.1016/j.cognition.2008.10.003

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2015). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015-0892-6

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196. doi: 10.3758/BF03206482

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*, 168–176. doi: 10.3758/BF03196154

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*, 558–584. doi: 10.1037/0033-295X.105.3.558

Willingham, D. B., Wells, L. A., Farrell, J. M., & Stemwedel, M. E. (2000). Implicit motor sequence learning is represented in response locations. *Memory & Cognition*, *28*, 366–375. doi: 10.3758/BF03198552

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or Differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33. doi: 10.1016/j.cogpsych.2015.04.002

Wills, A. J., Lea, S. E. G., Leaver, L. A., Osthaus, B., Ryan, C. M. E., Suret, M. B., ... Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: I. Unidimensional classification does not necessarily imply analytic processing; evidence from pigeons (Columba livia), squirrels (Sciurus carolinensis), and humans (Homo sapiens). *Journal of Comparative Psychology*, *123*, 391–405. doi: 10.1037/a0016216

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *The Quarterly Journal of Experimental Psychology*, *66*, 299–318. doi: 10.1080/17470218.2012.708349

Wills, A. J., O'Connell, G., Edmunds, C. E. R., & Inkster, A. B. (Accepted). Progress in modeling through distributed collaboration: Concepts, tools, and category-learning examples. *Psychology of Learning and Motivation*.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations

of formal models of categorization. *Psychological Bulletin*, *138*, 102–125. doi: 10.1037/a0025715

Yeates, F., Jones, F., Wills, A., Aitken, M., & McLaren, I. (2013). Implicit learning: A demonstration and a revision of a novel SRT paradigm. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (Vol. 1, pp. 3829–3834). Austin, TX: Cognitive Science Society.

Yeates, F., Jones, F. W., Wills, A. J., Aitken, M. R. F., & McLaren, I. P. L. (2012). Implicit learning: A demonstration and a revision of a novel SRT paradigm. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2581–2586). Austin, TX: Cognitive Science Society.

Yeates, F., Wills, A. J., Jones, F. W., & McLaren, I. P. L. (2012). State-Trace Analysis of Sequence Learning by Recurrent Networks. In *Proceedings of the 34th annual conference of the cognitive science society* (Vol. 1, pp. 2581–2586).

Yeates, F., Wills, A. J., Jones, F. W., & Mclaren, I. P. L. (2015). State-trace analysis: Dissociable processes in a connectionist network? *Cognitive Science*, *39*, 1047–1061. doi: 10.1111/cogs.12185

Zaki, S. R., & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: Evidence for multiple systems or task complexity? *Memory & Cognition*, *42*, 508–524. doi: 10.3758/s13421-013-0375-9

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398. doi: 10.3758/BF03193416

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
Vol. 68, No. 6, 1203–1222, http://dx.doi.org/10.1080/17470218.2014.978875

# Feedback can be superior to observational training for both rule-based and information-integration category structures

C. E. R. Edmunds[1], Fraser Milton[2], and Andy J. Wills[1]

[1]School of Psychology, University of Plymouth, Plymouth, UK
[2]Department of Psychology, University of Exeter, Exeter, UK

*(Received 17 October 2013; accepted 5 July 2014; first published online 9 January 2015)*



The effects of two different types of training on rule-based and information-integration category learning were investigated in two experiments. In observational training, a category label is presented, followed by an example of that category and the participant's response. In feedback training, the stimulus is presented, and the participant assigns it to a category and then receives feedback about the accuracy of that decision. Ashby, Maddox, and Bohil (2002. Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30, 666–677) reported that feedback training was superior to observational training when learning information-integration category structures, but that training type had little effect on the acquisition of rule-based category structures. These results were argued to support the COVIS (competition between verbal and implicit systems) dual-process account of category learning. However, a number of nonessential differences between their rule-based and information-integration conditions complicate interpretation of these findings. Experiment 1 controlled between-category structures for participant error rates, category separation, and the number of stimulus dimensions relevant to the categorization. Under these more controlled conditions, rule-based and information-integration category structures both benefited from feedback training to a similar degree. Experiment 2 maintained this difference in training type when learning a rule-based category that had otherwise been matched, in terms of category overlap and overall performance, with the rule-based categories used in Ashby et al. These results indicate that differences in dimensionality between the category structures in Ashby et al. is a more likely explanation for the interaction between training type and category structure than the dual-system explanation that they offered.

*Keywords*: Competition between verbal and implicit systems; COVIS; Categorization; Implicit; Explicit; Feedback.

Ashby and Maddox (2011) stated that many researchers now assume that multiple systems are involved in category learning. To the extent that this claim is accurate, it is down in no small part to the behavioural dissociations reported by Ashby, Maddox, and colleagues. These studies tend to find a differential effect of a manipulation on the learning of two types of category structure: rule-based and information-integration. Ashby and Maddox (2011) argue that these dissociations

Correspondence should be addressed to C. E. R. Edmunds, School of Psychology, Plymouth University, Drake Circus, Plymouth, PL4 8AA, UK. E-mail: charlotte.edmunds@plymouth.ac.uk

Our thanks go to Angus B. Inkster for his help coding the verbal reports.

are predicted by one particular dual-system model of category learning, COVIS (competition between verbal and implicit systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Paul, & Maddox, 2011), which assumes the existence of two competing systems of category learning. The strength of the case for COVIS is, of course, not a function of the number of dissociations that have been reported, but rather of the number that prove to be reliable and valid. Indeed, there is a growing body of work that casts doubt on the validity or interpretation of a high proportion of these dissociations (e.g., Dunn, Newell, & Kalish, 2012; Newell, Dunn, & Kalish, 2010; Newell, Moore, Wills, & Milton, 2013; Stanton & Nosofsky, 2007, 2013). In light of this accumulation of critiques, it becomes particularly important to assess the remaining dissociations. In the current article, we report a reexamination of an influential dissociation reported by Ashby et al. (2002), which has not been previously reexamined.

Ashby et al. (2002) compared the effect of observational and feedback training on categorization performance. On each trial in observational training, participants were shown the correct category label, followed by the stimulus, and then made a classification response. In feedback training, participants were shown the stimulus, made a classification response, and then received feedback on the accuracy of that response. The stimuli were lines that varied in length and orientation. Two different category structures were considered: a unidimensional rule-based structure, such as Figure 1a, and a two-dimensional diagonal information-integration structure, such as Figure 1b. Ashby et al. found that participants' performance in the unidimensional rule conditions were similar regardless of training type, whereas participants in the information-integration conditions were less accurate with observational training than those with feedback training. They argued that these findings support the COVIS (competition between verbal and implicit systems) model of category learning (Ashby et al., 1998, 2011).

Ashby et al.'s (2002) dissociation is predicted by COVIS because the model assumes that rule-based and information-integration category structures are most effectively learned via dissociable neural systems that utilize feedback differently (Ashby et al., 1998). The verbal system relies on explicit, logical reasoning and excels at learning rule-based categories by testing simple verbal rules such as "short lines belong to Category A, and long lines belong to Category B", such as Figure 1a, or conjunctive rules such as "large, horizontal lines belong to Category A, otherwise they belong to Category B", illustrated in Figure 1c. The verbal system operates via a process of hypothesis generation and testing that utilizes working memory to maintain representations of the stimulus and the current rule long enough to learn regardless of the order in which the information is presented (Ashby et al., 2002). Consequently, as found by Ashby et al., COVIS predicts that training type should have little effect on the learning of rule-based categories. In contrast, the implicit system integrates information from the multiple stimulus

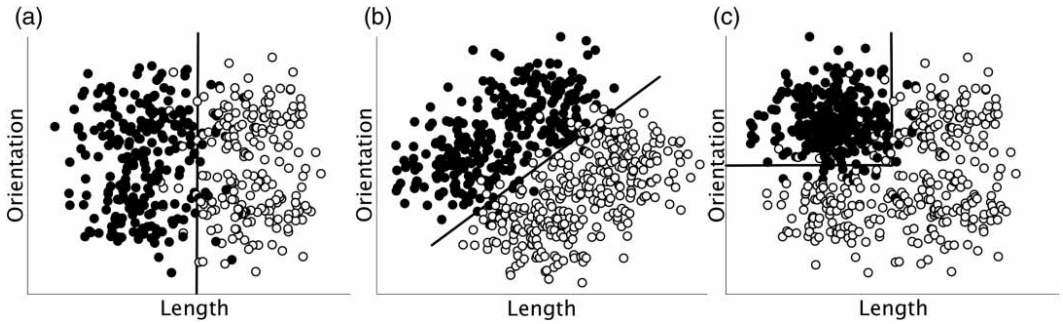


**Figure 1**. *Stimulus space representations of (a) a unidimensional category structure, (b) a diagonal or information-integration category structure, and (c) a conjunction category structure. Filled circles represent Category A, and unfilled circles represent Category B.*

dimensions predecisionally and associates this representation with a particular motor response. The implicit system is proposed to be responsible for learning "information-integration" category structures, illustrated in Figure 1b, where the perceptual boundary between the categories is difficult or impossible to describe verbally and therefore cannot be optimally learned by the verbal system (Ashby et al., 1998). The implicit system is hypothesized to be sensitive to how feedback is presented. It relies on unexpected reward to learn, so should learn more effectively when feedback follows a response than when the category label precedes the response (Ashby & Maddox, 2003). This means that COVIS predicts, as found by Ashby et al. (2002), that learning of information-integration categories should be impaired with observational training relative to feedback training.

In terms of COVIS, the critical difference between the rule-based and information-integration categories is that the former structure is readily verbalizable whereas the latter is not (Ashby et al., 1998). This is because verbalizability determines which system is responsible for optimum responding. Therefore, an ideal test of COVIS's predictions about the effect of training type on category learning should vary verbalizability while holding other potential confounds constant. However, Ashby et al.'s (2002) study contained three superfluous factors that varied between the rule-based and information-integration categories. First, the number of dimensions required to accurately learn each category varied: The information-integration structure required participants to utilize both stimulus dimensions, whereas the rule-based category structures only required one. Single-dimension classification has been shown to sometimes require less cognitive resources (as indexed by the effects of concurrent load and time pressure) than multidimension classification (Milton, Longmore, & Wills, 2008; Wills, Milton, Longmore, Hester, & Robinson, 2013). Therefore, training type may be less critical in the rule-based conditions than the information-integration conditions because it is a less demanding category structure.

Second, participants in Ashby et al.'s (2002) first experiment made very few errors in the rule-based conditions, but rather more in the information-integration conditions, raising the possibility that the observed dissociation was the result of a ceiling effect. Ashby et al. partially addressed this possibility by running a second study in which the rule-based structure was made harder to learn by reducing the between-category separation. Although the overall performance of participants decreased, there was still no statistically significant difference between observational and feedback training for rule-based categories under these conditions, supporting Ashby et al.'s interpretation. That being said, performance was marginally better with feedback training than with observational training. In addition, difficulty was only increased for one of the two combined counterbalance conditions, and there were only five participants per condition. Thus, the lack of a significant difference in Ashby et al.'s Experiment 2 might be attributable to a lack of statistical power.

Third, in both of Ashby et al.'s (2002) experiments, the rule-based structures had lower category separation than the information-integration structures. Category separation is the mean distance between category items as plotted in stimulus space, divided by the within-category variance along the direction of the comparison. Given that differences in category separation were shown by Stanton and Nosofsky (2007) to be responsible for the dissociation in another paper purported to support COVIS (Maddox, Ashby, Ing, & Pickering, 2004), it seems important to control for this factor in future investigations of Ashby et al.'s (2002) dissociation.

Although it is difficult to simultaneously control all three of these factors (number of relevant dimensions, error rates, and category separation) while maintaining the essential difference in verbalizability, this goal has been achieved in other COVIS-related studies. Specifically, Filoteo, Lauritzen, and Maddox (2010), in their study of the effects of concurrent load on rule-based and information-integration category learning, employed the category structures illustrated in Figures 1b and 1c. Filoteo et al.'s rule-based structure is a conjunctive rule and so requires participants to be sensitive to both stimulus dimensions. Furthermore, Filoteo

et al.'s study establishes empirically that these rule-based and information-integration structures are well matched on participant error rates. They are also closely matched on category separation.

For these reasons, Experiment 1 reexamined the effect of feedback compared to observational training using the category structures utilized by Filoteo et al. (2010). For this experiment, COVIS predicts that feedback training should be superior to observational training for the information-integration structure, but that training type should matter relatively little for the rule-based structure. However, Ashby et al.'s (2002) data are also consistent with the hypothesis that feedback is superior to observation for both rule-based and information-integration category structures. This is because the dissociation observed by Ashby et al. may be due to one or more of the superfluous factors for which they did not control (participant errors, category separability, problem dimensionality). Under this latter hypothesis, the current experiment should show a similar feedback advantage for both category structures, because these superfluous factors have been better controlled.

In addition to an examination of response accuracy, we also asked participants, at the end of the experiment, to describe their classification strategies. Not only does previous evidence indicate that reported strategy use can be informative when comparing the effect of feedback and observational training on a probabilistic category learning task (Newell, Lagnado, & Shanks, 2007), but it can also directly assesses whether participants can verbalize the category structure. If the rule-based, conjunction category structure is more verbalizable than the information-integration category structure, then participants should be more successful at describing the underlying structure in the rule-based condition than the information-integration condition. Also, the use of model-based analysis of participants' responses, based around general recognition theory (GRT; Ashby & Gott, 1988), is standard practice within experiments inspired by the COVIS model. Although we have some reservations about this procedure, we have presented these analyses to facilitate comparison with other work in this field.

# EXPERIMENT 1

## Method

### *Participants and apparatus*

A total of 80 participants (47 female) were recruited from the University of Exeter community and were not rewarded for their participation.

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a MacBookPro with a 15-inch screen.

### *Design*

The experiment had a 2 (category structure: rule-based, information-integration) × 2 (training type: observation, feedback) between-subjects, factorial design. A total of 20 participants were randomly assigned to each condition. Category learning was measured by the percentage of correct responses in each block.

### *Stimuli*

We used the same stimuli as that in the two-dimensional information-integration, Figure 1b, and rule-based, Figure 1c, conditions of Filoteo et al. (2010). Each stimulus was a single black line on a white background that varied on two dimensions: line length and orientation. In both conditions, maximum accuracy was 95% as 5% of the stimuli overlapped the optimal category boundary.

### *Procedure*

Participants in all conditions were informed that they would be shown a series of lines that varied in length and orientation, that their task was to assign the lines to either Category A or Category B, and that approximately half the lines were in each category. They were also told that at the beginning they might have to guess but by the end they should be able to reach high levels of accuracy. They were further informed of the structure of the experiment, the format of the trials, the position of feedback within the trial (which varied between conditions), and the response keys.

The experiment consisted of 10 blocks of 60 trials, with 600 trials in total. Participants assigned stimuli to either Category A (by pressing the "Z" key) or Category B (by pressing the "/" key). Starting with a training block, the blocks alternated between training and test. This was to provide a measure of performance during learning for both observational and feedback conditions as well as to facilitate comparison with Ashby et al. (2002). The training trials of the feedback learning conditions consisted of displaying the stimulus for 500 ms, followed by a blank screen for 500 ms, followed by a self-paced classification response. Finally the correct category label was displayed for 500 ms. In the observational learning condition, training trials consisted of first displaying the correct category label for 500 ms, followed by a blank screen for 500 ms, followed by the stimulus for 500 ms to which the participant made a self-paced response. The test trials in both feedback and observational training conditions included no information about the correct category assignment and consisted of a stimulus displayed for 500 ms followed by a self-paced response. The intertrial interval in all conditions was 500 ms.

At the end of the experiment, participants were presented with a questionnaire that asked them to describe whether they had a specific strategy when classifying the items and, if so, to describe it, using either words or pictures.

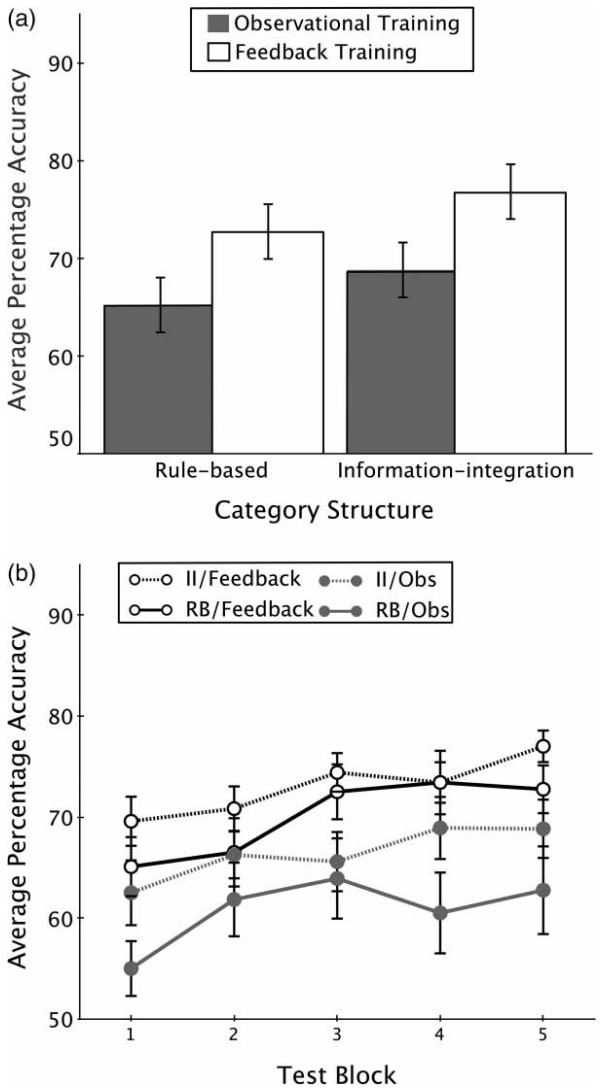


**Figure 2**. *(a) Percentage of correct responses by condition in the final (fifth) test block. (b) The average proportion of correct responses for each block in Experiment 1. II = information-integration; RB = rule-based; obs = observation. Error bars are one standard error.*

## Results

Following Ashby et al. (2002), analyses were conducted on the final test block of the data from all participants. Conducting the analyses across all test blocks led to the same conclusions, as did excluding participants failing to reach 50% on the final block (the analysis method and exclusion criterion applied by Filoteo et al., 2010). Figure 2 shows mean accuracy for each condition in just the last test block (Figure 2a), and across all test blocks (Figure 2b).

An analysis of variance (ANOVA) revealed a significant main effect of training type, $F(1, 76) = 7.68$, $\eta^2 = .09$, $p = .007$, but not of category structure, $F(1, 76) = 1.89$, $\eta^2 = .02$, $p = .175$. Hence, participants learned more in the feedback training condition than in the observational training condition when learning both rule-based and information-integration categories. The interaction between training type and category structure was also nonsignificant, $F(1, 76) = 0.058$, $\eta^2 = .00$, $p = .811$.

### *Bayesian analysis*

The standard statistical analyses above indicate that, unlike in Ashby et al. (2002), there appears

to be no difference between rule-based and information-integration categories in the effect of training type on learning. However, in null-hypothesis significance testing, nonsignificant results are ambiguous; they could either be due to insufficient statistical power or be due to the null hypothesis being correct (Dienes, 2011). As the interaction between training type and category structure formed the basis of the conclusions drawn by Ashby et al. (2002), it is important to determine whether the reason the current study failed to find an effect was because it lacked power. One way of determining this is to calculate Bayes factors for the relevant comparisons (Dienes, 2011). Briefly, if the Bayes factor is over three then the experiment has found evidence for the experimental hypothesis, whereas if the Bayes factor is less than a third, the experiment finds evidence for the null hypothesis (Jeffreys, 1961). A Bayes factor of one indicates that the evidence is exactly neutral with respect to the experimental and null hypotheses (Dienes, 2011). Values between a third and three are typically interpreted as indicating that the experiment was not sensitive enough, and no conclusions can be drawn.

To calculate the Bayes factor for the interaction between category structure and learning type we followed the procedure recommended by Dienes (2011). This requires the expected average difference between the two differences to be specified. In Ashby et al. (2002), the observed mean difference of differences between the information-integration conditions in Experiment 1 and the rule-based conditions in Experiment 2 was approximately 15%, and we used this figure in our analyses. This cross-experimental difference that was used as the rule-based structure in Experiment 2 was better controlled for differences in overall error rates. Following the recommendations of Dienes (2011), we assumed a normal distribution around this mean with standard deviation of half the mean (i.e., 7.5, representing the experimental hypothesis that differences as small as zero are unlikely). These calculations result in a Bayes factor of 0.18. As the Bayes factor is less than a third, it indicates that the data provide support for the null hypothesis—that is, that there is no difference between rule-based and information-integration category learning in the effect of varying training type. These conclusions held even if the expected average difference between the rule-based and information-integration conditions was underestimated by up to a third of that reported by Ashby et al.

### *State-trace plot*

The analyses above indicate that there are no differences between the acquisition of rule-based and information-integration categories. However, these analyses do not consider the qualitative pattern of learning throughout the experiment. As the key conceptual claim of COVIS is that there are two mechanisms of learning, it could be argued that these analyses have failed to identify multiple systems only because the difference in learning between training types just happened to be the same for rule-based and information-integration learning by the end of training. To examine the validity of this claim, we used state-trace analysis (Bamber, 1979; Loftus, Oberg, & Dillon, 2004), which has previously been used with great success on this type of category learning data (Dunn et al., 2012; Newell et al., 2010).

State-trace analysis is an alternative to dissociation logic that allows experimenters to determine whether multiple systems are required to explain an experimental result. This is accomplished by drawing a state-trace plot. To do this, two dependent variables, in this case performance on the rule-based and information-integration category structures, are plotted on the $x$ and $y$ axes. Then, a trace is plotted for each training type condition, with each point being the accuracy from each test block. The state-trace plot is then inspected to determine whether the traces are consistent with a single- or multiple-system account. If the two traces overlap to form a single monotonic function, then there is an absence of evidence that a multiple-process account is required to explain the observations. If the traces form two monotonic functions, then this is often interpreted as being more supportive of a multiprocess account, although the question of what the term "multiprocess" means in the context of state-trace analysis has been the topic of recent debate (Dunn, Kalish, &

Newell, 2014). In brief, both Yeates, Wills, Jones, and McLaren (2012, in press) and Ashby (in press) have identified situations where models typically considered to be single-system accounts can produce two functions on a state-trace plot through variation in a single parameter (specifically, attention weight in the generalized context model, Nosofsky, 1986, and learning rate in the simple recurrent network, Elman, 1990).

From visually inspecting Figure 3, the data from the current experiment form a single monotonic curve. This suggests an absence of evidence that a multiprocess account such as COVIS is required to account for the current results. However, it is worth noting that to conclusively infer this, the plot should be statistically tested for a significant departure from monotonicity.

### *Model-based analyses*

The COVIS-based predictions for this data set (see introduction) are contingent on the assumption that the category type manipulation corresponds to a change in the learning system that controls responding. Practically, this means that there should be more people using the verbal system in the rule-based category conditions than in the information-integration conditions, and vice versa for the implicit system. Experimental studies within the COVIS framework utilize model-based analysis constructed from GRT (Ashby & Gott, 1988) to examine this assumption. For each participant, this analysis determines the optimum decision boundary in stimulus space that separates the stimuli judged by each participant to be in Category A from those in Category B. Each participant is then assigned a strategy type, such as unidimensional, on the basis of characteristics of their optimum boundary. The assumption that the category type manipulation has resulted in a change of category learning system is argued to be valid if more participants are using the optimum decision bound for the category structure that they have been assigned to, such as a diagonal decision boundary in the information-integration conditions, than are using that strategy in the inappropriate category structure, such as a diagonal decision boundary in the rule-based conditions.

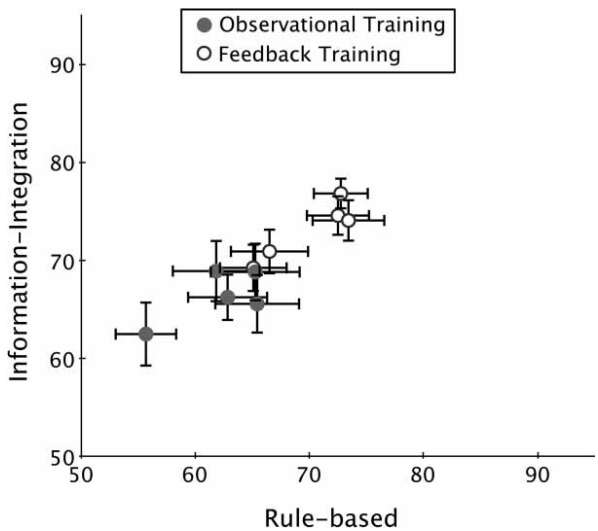


**Figure 3**. *State-trace plot with rule-based and information-integration performance on each block on the axes. Error bars are one standard error.*

The GRT-based analysis determines which of a predefined set of decision-boundary models best describes the classification each participant has produced. The set of models considered in this analysis were as follows:

The *unidimensional* models assume that the participant determines a criterion along one of the stimulus dimensions, either orientation or length. They then make a decision about the category membership of each stimulus by comparing the appropriate stimulus attribute with the criterion value. As an example, for length, this corresponds to a rule of the type: "Assign to Category A if the stimulus is long, or Category B if short". The unidimensional models have two parameters: the value of the criterion and the variance of internal (criterial and perceptual) noise.

The *conjunction* model assumes that the participants make two judgements, one for each stimulus dimension, and then combine these to make a judgement about category membership. The conjunction rule in the current analysis was of the type: "Assign to Category A if the stimulus is short and upright, otherwise assign to Category B". The conjunction model had three parameters: the two criterion values and internal noise.

The *general linear classifier* (GLC) model assumes that the decision boundary between the categories can be described by a straight line that can vary in

gradient and intercept. The unidimensional models are therefore special cases of the GLC model. The GLC model has three parameters: the intercept and slope of the decision bound, plus noise.

The *random* model assumes that participants are responding randomly; it has no parameters.

For each participant, the best fit of each of these models was calculated, and the best fitting model was selected using Akaike's information criterion (Akaike, 1974). The results from this analysis, which was performed using the grt package in the R environment (Matsuki, 2014), are reported in Table 1. Within the COVIS framework, the unidimensional and conjunction models are considered to represent explicit, rule-based strategies, while the GLC is considered to represent an implicit, information-integration strategy.

In ordinal terms, the results of this analysis are consistent with the intended effects of the experimental manipulation, as seen through the lens of the COVIS model and GRT-based model analysis. Specifically, the proportion of participants best fit by a conjunction model is higher in the rule-based condition than in the information-integration condition, and the proportion of participants best fit by the GLC model is higher in the information-integration condition than in the rule-based condition.

**Table 1.** *The proportion of participants in each condition according to the model-based analysis in Experiment 1*

| | *Strategies* | | | | |
|---|---|---|---|---|---|
| *Condition* | *CJ* | *UDO* | *UDL* | *GLC* | *RND* |
| RB-FB | .35 | .3 | .15 | .15 | .05 |
| RB-Obs | .4 | .2 | .1 | .2 | .1 |
| RB overall | .375 | .25 | .125 | .175 | .075 |
| II-FB | .15 | .45 | .1 | .3 | 0 |
| II-Obs | .3 | .3 | .05 | .35 | 0 |
| II overall | .225 | .375 | .075 | .325 | 0 |

*Note:* Conditions: RB-FB = rule-based/feedback; RB-Obs = rule-based/observation; II-FB = information-integration/feedback; II-Obs = information-integration/observation condition. Strategies: CJ = conjunction; UDO = unidimensional strategy based on orientation; UDL = unidimensional strategy based on length; GLC = general linear classifier; RND = random.

It is perhaps not particularly surprising that some participants are best fit by a unidimensional model, as a single-dimension strategy can optimally achieve approximately 75% accuracy in both the rule-based and the information-integration conditions. From a COVIS perspective, it is not particularly problematic if some participants in the rule-based condition are in fact employing a unidimensional rule, as this is still a rule-based strategy and readily verbalizable. It is potentially more problematic from a COVIS perspective that there are a reasonable proportion of participants best fit by unidimensional models in the information-integration condition, potentially implying the presence of significant rule-based responding in these conditions. A similar result was observed in Ashby et al. (2002), although the proportion is higher in the current study. The presence of unidimensional responders in an information-integration condition is typically accommodated within COVIS by assuming that some participants have not yet transitioned from the explicit system to the implicit system. The lower proportion of participants best fit by unidimensional models in Ashby et al. (2002) may be due to the fact that Ashby et al., in their modelling of their information-integration condition, constrained the GLC model to have the gradient and intercept defined by the category structure. This constrained version of the model has just one parameter, while the unconstrained version that we employed has three parameters. In an AIC model-selection procedure, reducing the number of free parameters of a model will, other things being equal, increase the proportion of participants best fit by that model. Somewhat surprisingly, Ashby et al. state that they employed the unconstrained version of the GLC in their fits of their rule-based condition. This difference in fitting procedure between experimental conditions seems odd and may have contributed to the higher proportion of unidimensional classifiers in their rule-based conditions than in their information-integration conditions.

In summary, the model-based procedures that are standard in this field broadly support the supposition that participants in the rule-based conditions classify the stimuli differently to participants in the

information-integration conditions. The fact that a conjunction model best fits more participants in the rule-based condition than in the information-integration condition, and a GLC model best fits more participants in the information-integration condition than in the rule-based condition, is broadly consistent with the predictions of the COVIS model. Of course, what is not consistent with the COVIS model is that, despite these differences, there is no difference in the size of the feedback advantage in the rule-based and information-integration conditions.

Although seldom reported within the COVIS literature, it is also informative to look at the performance of the best fitting model relative to the performance of the competing models. If the winning model performs much better than its competitors, we can be fairly confident that this model provides the best description of the participant's behaviour, from among the prespecified alternatives. On the other hand, if the competing models perform almost as well as the winning model, our confidence that the winning model provides the best description should probably be lower.

One principled way of evaluating the model-based analysis is by calculating the normalized probability that a conjunction model is preferred to the GLC for each participant (or vice versa). This is done by calculating the Akaike weight, $w_i$ (AIC), for each model for each participant (Wagenmakers & Farrell, 2004). This is defined as the probability that model $i$ is the best, in terms of minimizing the Akaike information criterion, given the data and the set of competing models. From the Akaike weights, the normalized probability that model $i$ is to be preferred over model $j$ is calculated using

$$\frac{w_i(\mathrm{AIC})}{w_i(\mathrm{AIC}) + w_j(\mathrm{AIC})} \tag{1}$$

where $w_i$ (AIC) and $w_j$ (AIC) are the Akaike weights for models $i$ and $j$, respectively.

For the rule-based category structure conditions the probability of the "best" model being a conjunction, rather than the GLC, is .635 in the feedback training condition and .668 in the observational training condition. This provides additional support that participants are genuinely using orthogonal decision boundaries to make decisions. However, for the information-integration category structures the probability of the best model being the GLC, rather than a conjunction, is much lower: .297 for the feedback training condition and .382 for the observational training condition. Clearly, confidence in the results of GRT-based model fitting in the information-integration conditions should be low. We would be interested to see comparable information for Ashby et al. (2002), or any other COVIS-relevant study, and suggest this or a similar measure be included in future research.

### *Verbal report analysis*

An alternative explanation of these findings from within the COVIS framework might be that the majority of participants in both the rule-based and information-integration category structure conditions were using the implicit system. It is possible that by increasing the number of relevant dimensions in the rule-based structure, participants found this too difficult and so resorted to using the implicit system. To investigate this possibility we examined the strategies reported by participants as summarized in Table 2.

The verbal reports were independently coded by two of the authors (C.E.R.E. and A.J.W.), and any discrepancies that were not due to human error were easily resolved through discussion. First, each verbal report was examined to determine whether the participant had reported an explicit categorization strategy or not. The inter-rater reliability for this was perfect, $\kappa = 1$, $p < .001$. Second, the available strategy descriptions were sorted into groups of three main kinds: unidimensional, two-dimensional, and miscellaneous.

Participants were placed in the *unidimensional length* or *unidimensional orientation* groups if they described categorizing stimuli based solely on line length or line orientation, respectively.

Participants were placed in the *conjunction* group if they used both stimulus dimensions and described categorizing stimuli using a logical conjunction rule such as "short, upright lines

**Table 2.** *The proportion of participants in each condition that reported using each strategy in Experiment 1*

| | *Strategies* | | | | | |
|---|---|---|---|---|---|---|
| *Condition* | *CJ* | *2D* | *UD* | *II/OS* | *Other* | *None* |
| RB-FB | .45 | .05 | .1 | 0 | .1 | .2 |
| RB-Obs | .2 | .15 | .2 | 0 | .2 | .25 |
| RB overall | .325 | .1 | .15 | 0 | .15 | .225 |
| II-FB | .4 | .25 | .25 | 0 | .15 | 0 |
| II-Obs | .2 | .15 | .35 | 0 | .05 | .25 |
| II overall | .3 | .15 | .3 | 0 | .025 | .125 |

*Note:* Conditions: RB-FB = rule-based/feedback; RB-Obs = rule-based/observation; II-FB = information-integration/feedback; II-Obs = information-integration/observation condition. Strategies: CJ = conjunction; 2D = generic two-dimensional; UD = unidimensional; II/OS = information-integration or overall similarity.

were in Category A, otherwise they were in Category B".

Participants were placed in the *information-integration* group if they described attempting to make the stimulus dimensions commensurable, such as "Stimuli for which the line was longer than it was upright should be assigned to Category A" or if they said anything that could be reasonably interpreted as a statement that they based their classification on overall similarity. Note that overall similarity descriptions are commonly found in other studies, not within the COVIS-framework, in which we have elicited verbal reports (e.g., Wills et al., 2013).

Participants were placed in the *two-dimensional* group if they described using both stimulus dimensions but with descriptions that were too unclear to be assigned to more specific categories.

All remaining participants were assigned to the *other* group, which included participants whose descriptions were too vague to be assigned to another group.

Inter-rater reliability for strategy assignment was high, $\kappa = .813$, $p < .001$, with the majority of discrepancies appearing to be due to human error in applying the strategy definitions, rather than any inherent ambiguity in the definitions themselves (as all discrepancies were rapidly resolved by reference to the strategy descriptions). There were no significant differences between all conditions in the number of participants who did not report a strategy, $\chi^2(1) = 0.12$, $p = .730$. With respect to the types of strategy reported, there are very different patterns of responding between the rule-based and information-integration category structure conditions. For the rule-based conditions, although there is clearly some variability, the modal strategy correctly described the conjunction structure. In addition, none of the participants in these conditions reported using an overall similarity or information-integration strategy, and only 20.1% reported using unidimensional strategies.

In contrast, no participant in the information-integration category structure conditions reported any strategy that could be interpreted as describing the structure of the information-integration category that they had been presented. In these conditions, participants were equally likely to report a unidimensional strategy as they were to report a conjunction rule, although strategies employing both dimensions were the majority indicating a sensitivity to the fact that both dimensions were relevant. This summary is supported by the fact that the number of participants in the rule-based category conditions who reported the optimal strategy for the categorization problem that they had been presented (44.8% of the people who reported strategies) was significantly different from those in the information-integration conditions who identified the correct strategy (0% of the people who reported strategies), $\chi^2(1) = 15.20$, $p < .001$.

In sum, although participants found neither category structure trivial to verbalize, participants in the rule-based category structure conditions were more able to verbalize the underlying category structure than those in the information-integration conditions. Thus, these analyses largely support the assertion that the rule-based category structure is more readily verbalizable than the information-integration category structure.

### *Comparing model-based analyses with verbal reports*

The model-based analyses and verbal reports used here are complementary approaches that both aim to determine how participants are completing the task. However, from the summaries of these analyses

above, it appears that they are partially inconsistent with each other. To examine the degree of correspondence between these approaches, we compared the strategy that each participant was assigned using the model-based analysis with the one they reported using after the experiment (Table 3).

As can be seen, for the rule-based strategies (unidimensional, two-dimensional, and conjunction) the model-based analyses and verbal reports match reasonably well. This is not the case for the GLC and reports of implicit or overall similarity responding; all participants that were assigned to the GLC strategy in the model-based analysis reported using an explicit rule-based strategy. One possible explanation for this disparity is that participants were using an implicit, GLC-based, strategy but were unable to describe it correctly. Although this may be unsurprising given that it is implicit, it seems unlikely given that in previous, different, but related, work participants were able to report this type of strategy (Wills et al., 2013). Alternatively, it may be that the GLC is more inclusive than the other models and so results in participants that are using a rule-based strategy being assigned to the GLC merely because they could not be assigned to another type of strategy. This latter hypothesis is supported by the Akaike weight for the GLC; this model wins by a much lower margin than the others (see "Model-based analysis" section).

## Discussion

Ashby et al. (2002) reported, as predicted by COVIS, that performance with feedback training was superior to that with observational training when learning an information-integration category structure, whereas for a unidimensional rule-based category they found that these training types resulted in comparable performance. In contrast, we found that learning performance in Experiment 1 was better with feedback training than with observational training to a similar degree for both category structures. The Bayesian analysis verifies that there is truly no difference in learning performance between the two category structures. This pattern of performance is not consistent with the claim that there are two systems of category learning that are differentially affected by training type. The state-trace analysis shown in Figure 3 also does not provide any evidence for a dual-system approach. It consists of a single, monotonically increasing curve, which is interpreted as evidence that performance in this experiment can be described by a single system of category learning.

COVIS could encompass the pattern of performance found in Experiment 1, if participants resorted to using the implicit system for both category structures. However, this hypothesis is not supported by the verbal report analysis that found that participants were equally likely to be able to report a strategy in all conditions, but that fewer participants were able to describe the optimal strategy in the information-integration conditions than in the rule-based conditions. Similarly, the model-based analysis indicates that the conjunction model best fits more participants in the rule-based condition than in the information-integration condition, and a GLC

**Table 3.** *Comparison of the models assigned to each participant in the model-based analysis with those that they reported using in Experiment 1*

| | *Verbal strategy reports* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Rule-based* | | | | *Information-integration* | | | |
| *Model-based strategies* | *UD* | *CJ* | *2D* | *II/OS* | *UD* | *CJ* | *2D* | *II/OS* |
| UD | 6 | 1 | 3 | 0 | 8 | 1 | 5 | 0 |
| CJ | 4 | 10 | 0 | 0 | 4 | 5 | 0 | 0 |
| GLC | 2 | 2 | 3 | 0 | 2 | 6 | 4 | 0 |

*Note:* UD = unidimensional; CJ = conjunction; GLC = general linear classifier; 2D = two-dimensional strategy; II/OS = either an information-integration or overall similarity strategy.

model best fits more participants in the information-integration condition than in the rule-based condition. Therefore, the results of Experiment 1 appear inconsistent with COVIS.

## EXPERIMENT 2

Ashby et al. (2002) found an interaction between training type and category structure. They argued that this pattern of results supported COVIS. However, Ashby et al. included several confounds in their design that complicate interpretation of their results: the number of stimulus dimensions relevant to categorization, category separation, and error rates. When these were controlled for in our Experiment 1, feedback training was superior to observational training when learning both rule-based and information-integration categories—a pattern of results not consistent with COVIS. The key difference between Experiment 1 and Ashby et al.'s findings is the appearance of a feedback training advantage for rule-based categories. Experiment 2 of the present paper aimed to determine which of the controlled for confounds might have resulted in this difference in the effects of training type.

The number of dimensions relevant to classification seemed to be the most likely cause of the difference between our Experiment 1 and Ashby et al. (2002). This is because Ashby et al. (2002) manipulated category separation and error rates in a second experiment and still did not find a statistically significant difference in performance due to training type. Therefore, in Experiment 2 of the current paper, to discriminate dimensionality from the other factors, the number of relevant dimensions in the category structure were maintained whilst category separation and error rates were varied. Category separation was increased. Error rates were reduced by scaling the length dimension to increase perceptual discriminability along that dimension and on each trial the stimulus, category label, and intertrial interval were increased to 1000 ms.

If increased error rates or reduced category separation are the cause of the difference in learning rule-based categories between our first experiment and Ashby et al. (2002) then the difference between training type should disappear in this experiment. However, if the locus of the difference is the number of relevant dimensions for the rule then the advantage for feedback training over observational training should remain.

### Method

#### *Participants and apparatus*

A total of 40 participants (10 male) were recruited from the Plymouth University paid pool and were paid £8 for their participation.

The experiment was run on a desktop computer on a 21.5″ screen using MATLAB 2012b with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

#### *Design*

The experiment had two between-subjects conditions (training type: observation, feedback). A total of 20 participants were randomly assigned to each condition. Category learning was measured by the percentage of correct responses in each test block.

#### *Stimuli*

This version of the experiment still utilized a conjunction category structure. However, the category structure was altered to make learning easier (Figure 4). To generate the category structure, four sets of points, 300 from the Category A distribution and 100 each from the other three, were randomly selected from bivariate normal distributions defined using the parameters listed in Table 4. Any points that were over 2.25 standard deviations away from the mean of the distribution in the direction of the category boundary were resampled. Then, as Experiment 1 indicated that the orientation of the line stimuli appeared more salient than line length to participants, the distribution was scaled so that the lines varied between 20 and 350 points in arbitrary units.

#### *Procedure*

The only change to the procedure of Experiment 1 was that intertrial interval and the duration of the stimulus and category label presentation were increased from 500 ms to 1000 ms.

After the experiment, participants again completed a questionnaire to identify the strategy they used. The format of this varied slightly from the one used in Experiment 1, based on our experience of coding the Experiment 1 responses and in an attempt to elicit clearer descriptions. They were asked to: "Imagine that another participant was asked to complete the experiment exactly as you did. What instructions would you give them so that they could exactly copy your pattern of responding? Please try to be as precise as possible."

## Results

Following Ashby et al. (2002) and Experiment 1, analyses were conducted on the final test block of the data from all participants. Conducting the analyses across all test blocks led to the same conclusions. No participant failed to reach 50% accuracy by the final test block. Figure 5 shows mean percentage accuracy in each condition for all test blocks.

Overall performance, as expected, was higher than that for the participants in the rule-based condition in Experiment 1. An ANOVA revealed a statistically significant effect of training type, $F(1, 38) = 4.61$, $\eta^2 = .108$, $p = .038$. Hence, as in Experiment 1, participants learned consistently more in the feedback training condition than in the observational training condition (see Figure 5).

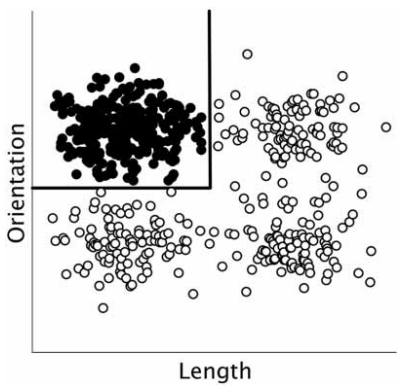


**Figure 4**. *The conjunction category structure used in Experiment 2. Filled circles represent Category A, and unfilled circles represent Category B.*

**Table 4.** *Parameters used to generate the initial stimulus distribution for Experiment 2*

| | *Parameters* | | | |
|---|---|---|---|---|
| *Category* | $\mu_I$ | $\mu_O$ | $\sigma_I$ | $\sigma_O$ |
| A | 100 | 200 | 20 | 20 |
| B | 100 | 100 | 20 | 20 |
| B | 200 | 100 | 20 | 20 |
| B | 200 | 200 | 20 | 20 |

*Note:* Each row describes a set of points in stimulus space generated by a bivariate normal distribution with means ($\mu_I$, $\mu_O$) and standard deviations $\sigma_I$ and $\sigma_O$ for the length and orientation dimensions, respectively.

### *Model-based analysis*

The proportions of participants using each model in each condition are in Table 5. From this we can see that the majority of participants in both conditions have been identified by the analysis as using either the correct conjunction strategy or another rule-based one. This supports the hypothesis that the participants are using an explicit, rule-based strategy. However, the proportions of participants in each condition that were assigned to the correct conjunction strategy are statistically different, $\chi^2(1) = 5.63$, $p = .018$. This indicates that participants were more successful at determining the underlying category structure in the feedback training condition than in the observational training condition.

We also looked at the performance of the best fitting model relative to the performance of the competing models in terms of the Akaike weights. For the participants in the feedback training condition the mean normalized probability of using a rule-based strategy compared to information-integration or random strategies is .931, whereas for the observational training conditions the normalized probability is .770. This demonstrates that, as expected, participants are most likely to use rule-based strategies in these rule-based conditions and that these strategies were clear winners.

### *Verbal reports*

The verbal reports were independently coded by one of the authors (C.E.R.E.) and an independent

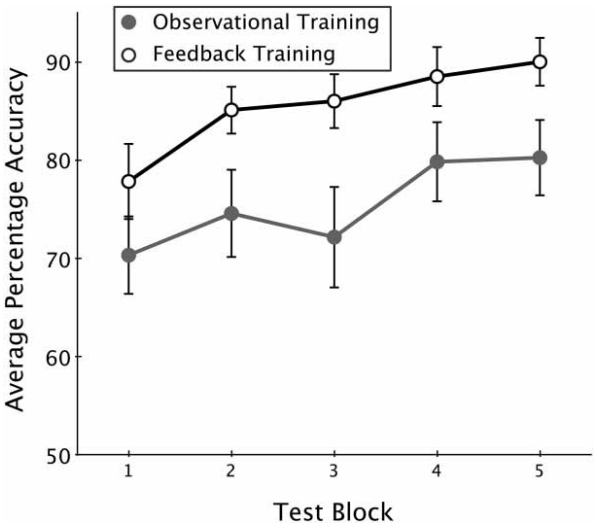


**Figure 5**. *The average proportion of correct responses for each block in Experiment 2. Error bars are one standard error.*

rater (A.B.I.). Any discrepancies that were not due to human error were easily resolved through discussion.

Inter-rater reliability for judging whether or not each participant reported a strategy was high, $\kappa = .844$, $p < .001$, whereas judgements as to which strategy they were reporting were reasonable, $\kappa = .595$, $p < .001$. The majority of discrepancies were due to different interpretations of how participants might be expected to describe a conjunction category structure. The coded strategies described by participants are shown in Table 6.

There were no significant differences between conditions in the number of participants who did not report a strategy, $\chi^2(1) = 0.36$, $p = .548$. There was also no significant difference between conditions in those who reported the correct conjunction category, $\chi^2(1) = 0.96$, $p = .327$. Therefore, not only were participants in both conditions capable of coming up with a strategy, but the majority were also able to correctly describe the category structure.

***Comparing model-based analyses with verbal reports***

As before, we also looked at the degree of correspondence between the verbal reports given by participants and the model that best fitted their responses as determined by the model-based analysis (Table 7).

In this experiment, the verbal reports matched the model-based analysis reasonably well; the majority of participants that reported using a conjunction strategy were also assigned to this in the model-based analysis. Furthermore, as might be expected in learning a rule-based category structure, no participants reported using implicit or overall similarity responding or were best described, in the model-based analysis, by the GLC model.

## Discussion

The key difference between Experiment 1 and Ashby et al. (2002) was the appearance of an advantage for feedback training over observational training when learning a rule-based category structure. Experiment 2 aimed to determine which of the factors that varied between these two experiments was responsible for this difference. To do this, Experiment 2 compared performance with feedback and observational training when learning a

**Table 5.** *Proportions of participants best described by each model according to the model-based analysis in Experiment 2*

| | *Strategies* | | | | |
|---|---|---|---|---|---|
| *Condition* | *CJ* | *UDR* | *UDL* | *GLC* | *RND* |
| Feedback | .9 | 0 | 0 | .05 | .05 |
| Observation | .65 | 0 | .1 | .05 | .2 |

*Note:* CJ = conjunction; UDO = unidimensional based on orientation; UDL = unidimensional based on length; GLC = general linear classifier; RND = random.

**Table 6.** *The strategies reported by each participant in Experiment 2*

| | *Strategies* | | | | | |
|---|---|---|---|---|---|---|
| *Condition* | *CJ* | *2D* | *UD* | *II/OS* | *Other* | *None* |
| RB-FB | .70 | .20 | .05 | 0 | 0 | .05 |
| RB-Obs | .55 | .20 | .15 | 0 | 0 | .10 |
| RB Overall | .625 | .20 | .10 | 0 | 0 | .075 |

*Note:* Conditions: RB-FB = Rule-based/feedback, RB-Obs = Rule-based/observation. Strategies: CJ = conjunction, 2D = generic two-dimensional, UD = unidimensional, II/OS = information-integration or overall similarity.

two-dimensional category, with reduced error rates and increased category separation compared with the category structure used in Experiment 1. Under these conditions, the advantage of feedback training over observational training remained. In addition, the model-based and verbal reports indicate that the majority of participants in both conditions were able to use and verbally describe a conjunction strategy. This indicates that the interaction between training type and category structure in Ashby et al. (2002) appears to be due to differences in dimensionality between the category structures.

## GENERAL DISCUSSION

Ashby et al. (2002) reported that feedback training was superior to observational training for an information-integration category structure, but that the two training types were comparable for a rule-based category structure. This dissociation has widely been taken as support for the COVIS dual-process theory of category learning (Ashby et al., 1998, 2011) and is the most cited, uncritiqued behavioural support for this model. According to the COVIS framework, the critical manipulation in Ashby et al. (2002) is that rule-based category structures are easily verbalizable, while information-integration categories are not and that this results in participants learning these two types of category structures using different category learning systems. These two systems incorporate feedback differently, therefore accounting for the Ashby et al. findings. However, there were several nonessential differences between the category structures used by Ashby et al., which casts doubt on whether verbalizability is the key factor in eliciting a differential effect of training type on learning performance.

In Experiment 1, we successfully maintained the between-category structure difference in verbalizability while matching them for (a) the number of relevant stimulus dimensions, (b) category separation, and (c) overall performance. We did this by combining the procedures of Ashby et al. (2002) with two-dimensional category structures adopted from more recent work in the COVIS framework (specifically Filoteo et al., 2010). Once these extraneous factors were controlled for, the category structure by training type interaction found by Ashby et al. did not appear; learning of both category structures was better with feedback training than with observational training. Experiment 2 also found a training type difference in learning the two-dimensional rule-based structure when this structure was broadly matched, in terms of category overlap and overall performance, with the rule-based structures used by Ashby et al. This indicates that the appearance of a differential effect of training type on rule-based learning in these experiments appears to be due to the two-dimensional nature of the conjunction structure; these experiments demonstrated an advantage for feedback training over observational training not only for information-integration categories, but also for two-dimensional rule-based categories.

**Table 7.** *Comparison of the models assigned to each participant in the model-based analysis with those that they reported using*

| | *Verbal strategy reports* | | | |
|---|---|---|---|---|
| *Model-based* | *UD* | *CJ* | *2D* | *II/OS* |
| UD | 0 | 1 | 2 | 0 |
| CJ | 2 | 22 | 6 | 0 |
| GLC | 0 | 0 | 0 | 0 |
| RND | 2 | 2 | 0 | 0 |

*Note:* UD = unidimensional; CJ = conjunction; GLC = general linear classifier; RND = random; 2D = two-dimensional strategy; II/OS = either an information-integration or overall similarity strategies.

### Alternative explanations

Our findings have implications for the COVIS theory of category learning because they are not predicted by COVIS. In Experiment 1, COVIS predicts a greater feedback advantage for the information-integration structure than the rule-based structure, but both conditions benefit from feedback training to a similar degree. In Experiment 2, COVIS does not predict a feedback advantage, yet one is observed. How, then, might the results

of both Ashby et al. (2002) and the current paper be explained?

First, we need to explain why feedback training is superior to observational training. Any theory that presumes learning is driven by prediction error (see e.g., Wills et al., 2009, for a review) should be able to accommodate this result because, in observational training, there is nothing to predict. The ALCOVE (attention learning covering map) model (Kruschke, 1992) is one of several possible category learning models in which learning is driven by prediction error, as is the striatal pattern classifier (Ashby & Waldron, 1999) that forms the basis of Ashby's explanation of why a feedback advantage is sometimes observed.

Second, we need to explain why a benefit of feedback training is sometimes not observed. One possibility is that such findings represent absence of evidence rather than evidence of absence. In Ashby et al.'s (2002) first experiment, performance on the harder, observational, training condition is close to ceiling, potentially obscuring the effect. In addition, Ashby et al. report a significant feedback advantage for the unidimensional category structure in the first test block (Ashby et al., 2002, p. 673), which smoothly reduces throughout training until it disappears in the final block (Ashby et al., 2002, Figure 3). Ashby et al.'s conclusions are based on the final block. In Ashby's second experiment, there is a numerical trend in the direction we predict, sample sizes are small, and only one of the two counterbalance conditions were below ceiling. Thus, one possibility is that feedback is always advantageous in rule-based category learning, but that some experiments fail to reveal this due to methodological issues.

Another possibility is that the feedback advantage is genuinely absent for single-dimension rule-based category structures, or at least much smaller than it is for multidimensional category structures (rule-based or otherwise). Although further research would be required to make this claim securely, it is interesting to speculate how such an effect might be explained if it were to be confirmed. One possibility is that the size of the feedback advantage is related to how effortful the classification is. Dimensional summation theory (Milton & Wills, 2004) predicts that single-dimension classification is less effortful than multidimensional classification, and this prediction has been supported in multiple studies (e.g., Milton et al., 2008; Wills et al., 2013).

In summary, COVIS predicts that there should be an interaction between training type and category structure, with a smaller difference between training types when learning a readily verbalizable category structure than when learning a hard-to-verbalize one. However, the available evidence (from both Ashby et al., 2002, and the current studies) indicates that the pattern of performance on these tasks might be better explained by an interaction of training type and the number of dimensions relevant to classification. Of course, these experiments have not completely disentangled verbalizability from dimensionality. In order to do this, one would have to examine the effect of training type on a unidimensional, difficult-to-verbalize category. This would be difficult as it is hard to conceive of a unidimensional category structure that would be hard to verbalize without redefining what is meant by a stimulus dimension.

More generally, although there is reasonable support for the idea that providing an opportunity for error improves learning (Grimaldi & Karpicke, 2012; Kornell, Hays, & Bjork, 2009; Potts & Shanks, 2014), such an effect is not always seen even in multidimensional category structures (Newell et al., 2007), and, in some memory tasks, the effect is even reversed (Haslam, Hodder, & Yates, 2011). Neither COVIS, nor our alternative explanation, fully captures these results. Further empirical work is required to clearly identify the conditions under which feedback training is superior to observational training.

## Dimensionality

As discussed above, it seems likely that it is the problem dimensionality, rather than the problem verbalizability, that drives the results of Ashby et al. (2002) and the current paper. The comparison of a unidimensional rule-based category structure

with a 45-degree rotation of that structure in stimulus space has formed the basis of a large number of experiments by Ashby and colleagues. The comparison is initially appealing, because the two structures are in various formal senses identical (e.g., an optimal classifier performs equally well on both structures), yet one is easy to verbalize while the other is hard to verbalize. However, the two structures are not matched on the number of psychological stimulus dimensions relevant to the classification. This raises the broader question of whether a failure to control problem dimensionality underlies other apparently COVIS-supporting dissociations.

A reviewer suggested that dimensionality is unlikely to be driving the difference of our results and those of Ashby et al. (2002) on the basis that pigeons find the two problems equally difficult (Smith et al., 2011), the implication being that if a nonverbal species finds these two problems equally difficult then it must be the verbalizability of the problems rather than their dimensionality that is important. However, even in nonverbal species, a necessary condition of a unidimensional problem being easier than a two-dimensional problem is that the stimulus dimensions are psychologically separable. Without separability, there is no meaningful psychological sense in which the two problems differ in dimensionality. Smith et al. (2011) provide no compelling evidence that their stimuli are separable for pigeons.

Another possible response to our claim that dimensionality is the critical factor is to point out that many of the more recent COVIS-supporting dissociations make use of a two-dimensional rule-based structure, thus equating problem dimensionality between rule-based and information-integration problems (e.g., Maddox, Bohil, & Ing, 2004; Maddox, Filoteo, Hejl, & Ing, 2004; Maddox, Filoteo, & Lauritzen, 2007; Maddox & Ing, 2005; Maddox, Love, Glass, & Filoteo, 2008; Zeithamova & Maddox, 2006), and dissociations, predicted by COVIS, still emerge. However, this evidence is not, perhaps, as compelling as it first appears, and in recent years it has attracted substantive critiques on a variety of bases from separate labs (e.g., Dunn et al., 2012; Newell et al., 2010, 2013; Stanton & Nosofsky, 2013; Zaki & Kleinschmidt, 2013). Our explanation is, therefore, entirely compatible with the existing evidence.

## Model-based analysis

Another interesting question raised by this research pertains to the limitations of the GRT informed model-based analysis which is ubiquitously used in analysing experiments within the COVIS framework. This model-based analysis aims to determine how participants are approaching the categorization task and from this make inferences as to which system is guiding responding. This model-based analysis is commonly interpreted by Ashby, Maddox, and colleagues to demonstrate that the category structure factor has successfully manipulated the learning system in control of responding if, for each category structure condition, more participants are assigned the correct strategy than the one appropriate for the other condition. The current work found this between-condition shift in strategies. However, the current work also utilized verbal reports and a state-trace analysis, which, although consistent with each other, are not consistent with the model-based analysis or its interpretation as supporting a dual-system approach. Visual inspection of the state-trace plot does not provide any evidence for multiple systems. Similarly, participants in all conditions were equally able to provide verbal reports. In addition, when examining the goodness-of-fit of each type of model in the model-based analysis using Akaike weights, there seems to be a disparity in the confidence the analysis places in the conjunction and GLC models that might also cast doubt on whether an actual switch between systems has taken place. This is obviously not the place for a detailed discussion and investigation of the conditions under which the model-based analysis is useful. However, future work might determine whether this type of model-based analysis is merely ineffective in this study, or whether it is more generally capturing something different from what was previously thought.

## Evidence for COVIS

This paper adds to the growing body of literature that has critiqued the experimental dissociations argued to support COVIS (Newell, Dunn, & Kalish, 2011). However, it is also important to note that there are a number of other dissociations that provide support for COVIS that have not yet been challenged. For example, switching response location part-way through training has been found to impact learning information-integration categories, while this manipulation does not affect rule-based category learning (e.g., Ashby, Maddox, Glass, O'Brien, & Filoteo, 2010). Deferring feedback has been found to have a similar impact on learning these two types of category structure (Smith et al., 2014). One might also point to imaging studies that show different neural substrates for rule-based and information-integration category learning (Ashby & Maddox, 2011, but see Milton & Pothos, 2011). Clearly, more work is needed to assess the strength of these and other dissociations taken to support COVIS.

## Conclusion

In summary, the current paper casts doubt on the interpretation of the dissociation found by Ashby et al. (2002). The current experiments demonstrated an advantage for feedback training over observational training not only for information-integration categories, but also for two-dimensional rule-based categories. Therefore, category structure dimensionality, rather than verbalizability, appears to be the key factor driving the appearance of an interaction between category structure and training type in the original study. This paper, therefore, adds to the growing literature (e.g., Dunn et al., 2012; Newell et al., 2010, 2013; Stanton & Nosofsky, 2007, 2013) that casts doubt on the validity or interpretation of the experimental evidence for the COVIS model of category learning.

## SUPPLEMENTAL MATERIAL

The underlying research materials for this article can be accessed at www.willslab.co.uk. The trial-level raw data for Experiment 1 are archived at www.willslab.co.uk/exe201201/ with md5 checksum 3a294887e14de59dc09bab76c27a9162.[1] The trial-level raw data for Experiment 2 are archived at www.willslab.co.uk/ply27/ with md5 checksum7beb8e5354548f9852a5412b8ac0dbd9.

## REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. Automatic control. *IEEE Transactions On*, *19*, 716–723.

Ashby, F. G. (in press). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems?. *Psychonomic Bulletin & Review*, 1–12.

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.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53.

Ashby, F. G., & Maddox, W. T. (2003). Dissociating explicit and procedual-learning based systems of perceptual category learning. *Behavioral Processes*, *66*, 309–332.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677.

Ashby, F. G., Maddox, W. T., Glass, B. D., O'Brien, J. B., & Filoteo, J. V. (2010). Category label and response location shifts in category learning. *Psychological Research*, *74*, 219–236.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.),

[1]Publication of an MD5 checksum allows the reader to independently confirm that the raw data in the archive are unchanged.

*Formal approaches in categorization* (pp. 1–13). New York: Cambridge University Press.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*, 363–378.

Bamber, D. (1979). State-trace analysis: A methods of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, *6*, 274–290.

Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be anappropriate tool for assessing the number of cognitive systems: A reply to Ashby (2014). *Psychonomic Bulletin & Review* (Advanced online publication).

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 840–859.

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

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*, 415–423.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding?. *Memory & Cognition*, *40*, 505–513.

Haslam, C., Hodder, K. I., & Yates, P. J. (2011). Errorless learning and spaced retrieval: How do these methods fare in healthy and clinical populations?. *Journal of Clinical and Experimental Neuropsychology*, *33*, 432–447.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford: Oxford University Press.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.

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

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*, 835–863.

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., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning–based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*, 945–952.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 227–245.

Maddox, W. T., Filoteo, J. V., & Lauritzen, J. S. (2007). Within-category discontinuity interacts with verbal rule complexity in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 197–218.

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.

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.

Matsuki, K. (2014). grt: General recognition theory. r package version 0.2. Retrieved from http://CRAN.R-project.org/package=grt

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

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*, 1326–1336.

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

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*, 563–581.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B.H. Ross (Ed), *The psychology of learning & motivation* (Vol 54, pp. 167–215). Oxford: Elsevier.

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.

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves implicit perceptual categorization: A comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*, 386–389.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667.

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18*, 414–421.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447–457.

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.

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1174–1191.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Wills, A. J., Lea, S. E. G., Leaver, L. A., Osthaus, B., Ryan, C. M. E., Suret, M. B., … Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: I. Unidimensional classification does not necessarily imply analytic processing; evidence from pigeons (Columba livia), squirrels (Sciurus carolinensis), and humans (Homo sapiens). *Journal of Comparative Psychology*, *123*, 391–405.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification?. *The Quarterly Journal of Experimental Psychology*, *66*, 299–318.

Yeates, F., Wills, A. J., Jones, F., & McLaren, I. (2012). State-trace analysis of sequence learning by recurrent networks. In *Proceedings of the 34th annual conference of the cognitive science society* (pp. 1–6).

Yeates, F., Wills, A. J., Jones, F. W., & McLaren, I. P. L. (in press). State trace analysis: Dissociable processes in a connectionist network?. *Cognitive Science*.

Zaki, S. R., & Kleinschmidt, D. F. (2013). Procedural memory effects in categorization: Evidence for multiple systems or task complexity?. *Memory & Cognition*, *42*, 508–524.

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

# Memory for exemplars in category learning

**C. E. R. Edmunds (ceredmunds@gmail.com)**
School of Psychology, Plymouth University
Plymouth, PL4 8AA, UK

**Andy J. Wills (andy.wills@plymouth.ac.uk)**
School of Psychology, Plymouth University
Plymouth, PL4 8AA, UK

**Fraser N. Milton (F.N.Milton@exeter.ac.uk)**
School of Psychology, University of Exeter
Exeter, EX4 4QG, UK

**Abstract**

Some argue that category learning is mediated by two competing learning systems: one explicit, one implicit (Ashby et al., 1998). These systems are hypothesised to be responsible for learning rule-based and information-integration category structures respectively. However, little experimental work has directly investigated whether people are conscious of category knowledge supposedly learned by the implicit system. Here we report one experiment that directly compared explicit recognition memory for exemplars between these two category structures. Contrary to the predictions of the dual-systems approach, we found preliminary evidence of superior exemplar memory after information-integration category learning compared to rule-based learning. This result is consistent with the hypothesis that participants learn information-integration category structures by using complex rules.

**Keywords:** category learning; memory; dual-systems; recognition;

One approach to categorization assumes that generalisation from past experiences to novel ones is mediated by two competing systems: one explicit and one implicit. The COVIS (COmpetition between Verbal and Implicit Systems) model is a popular instantiation of this approach (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Paul, & Maddox, 2011). In this model, the explicit, verbal system is described as learning by testing hypotheses using working memory. This system is assumed to optimally learn rule-based category category structures that can be easily verbalised such as the unidimensional structure shown in Figure 1A. In contrast, the implicit system is described as using procedural learning to associate areas of stimulus space with a motor response. This system is assumed to optimally learn category structures that cannot be easily verbalised such as the information-integration category structure shown in Figure 1B. Critically, the implicit system is assumed to "produce category knowledge that is opaque to declarative consciousness" (p.1, Smith et al., 2015).

Whether COVIS, and more broadly a dual-systems approach, adequately explains the processes of category learning is still a matter for debate. Proponents of COVIS argue that the case is closed, that the evidence for dual-systems approaches is overwhelming and that the field should move on to more interesting questions, such as those concerning the exact nature of the systems and how they interact with each other (Ashby & Maddox, 2011). In support of this view there is a large quantity of evidence that has been used to support COVIS (for a review see Ashby & Maddox, 2011). However, much of this evidence has been questioned (Edmunds, Milton, & Wills, 2015; Newell, Dunn, & Kalish, 2011; Stanton & Nosofsky, 2007; Zaki & Kleinschmidt, 2014). Also, despite the volume of studies, there is very little experimental work that directly investigates the key theoretical assumption that the learning of the implicit system is not available to consciousness. Instead, the focus has been on demonstrating that information-integration category structures are learned procedurally or demonstrating that learning of rule-based and information-integration categories are dissociable using experimental, neuropsychological or neuroscientific methods (Ashby & Maddox, 2005, 2011; Price, Filoteo, & Maddox, 2009). Therefore, the claim that the case is closed may be premature.

In the current study, we directly examine whether participants have conscious access to information about the information-integration categories they have learned. COVIS predicts that they do not, but some recent behavioral and neuroimaging evidence suggests otherwise. Behaviorally, Edmunds et al. (2015) found that the vast majority of participants were able to report a clear explicit strategy after training, regardless of whether they had been learning an information-integration or rule-based category structure. This was despite having met the criteria usually used in the COVIS literature to check that participants in the information-integration category structure condition are using the implicit system. This check uses a model-based strategy analysis inspired by General Recognition Theory (GRT; Ashby & Gott, 1988), a multidimensional version of signal detection theory. The failure of the model-based strategy analysis here may be because its output depends strongly on the set of strategies the modeller chooses to use, with the estimated proportion of "implicit" responders reducing substantially if more complex rule-based models are included (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015).

Turning to neuroimaging evidence, in a recent study from our lab we found greater activation of the medial temporal lobe in information-integration category learning, rela-

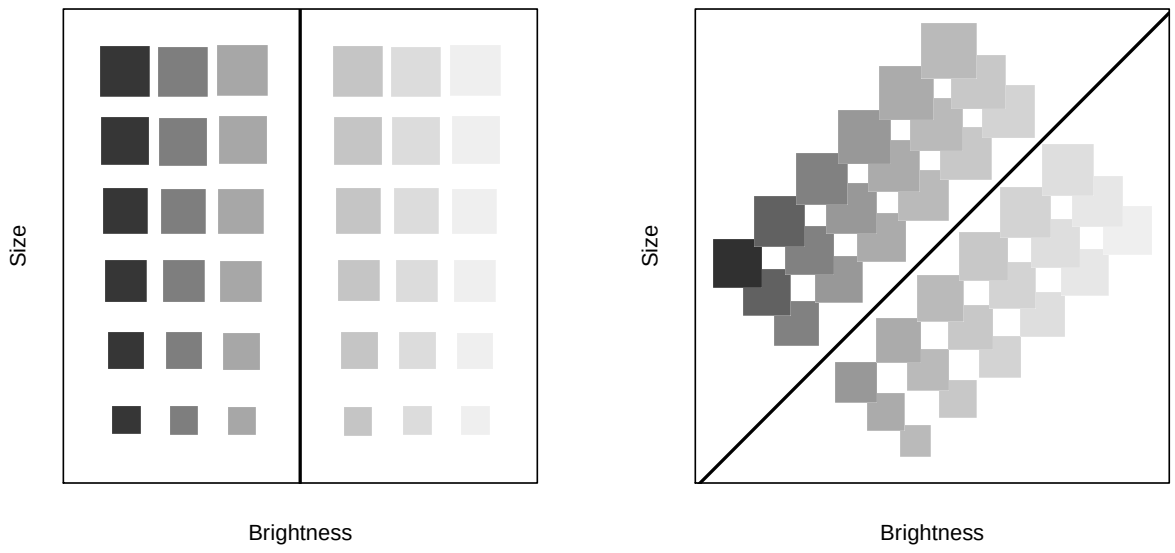


Figure 1: a) The unidimensional category structure with a rule based on size. b) The information-integration category structure with a positive category boundary.

tive to a well-matched rule-based category learning condition (Carpenter, Wills, Benattayallah, & Milton, in press). The medial temporal lobe has long been considered critical for explicit memory. Therefore, Carpenter et al.'s results suggest that information-integration category learning involves explicit memory processes to a greater extent than rule-based category learning.

One hypothesis that explains these two findings is that both rule-based and information-integration category structures are learned through the application of simple rules but that, for information-integration ones, those rules are supplemented by explicit memory of specific examples. For example, participants may store the examples that are exceptions to the simple rule. This hypothesis is also consistent with other evidence that found that participants who use a rule-plus-exception strategy have greater recognition memory for the exceptions to a simple category rule (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). As participants learning information-integration categories would have to remember more exceptions than those learning a rule-based structure, we would also expect greater recognition performance for information-integration learners.

However, this evidence is not conclusive. Neither Edmunds et al. (2015) nor Carpenter et al. (in press) directly measure explicit access to category knowledge. Further, some of the results of Carpenter et al. are at variance with a previous neuroimaging study of rule-based and information-integration category learning (e.g. Nomura et al., 2007). Carpenter et al. argue that the differences between their study and that of Nomura et al. are due to methodological problems with the Nomura et al. study. This argument is supported by the fact that Carpenter et al.'s results are broadly in line with the only other closely-related neuroimaging study (Milton & Pothos, 2011). Nevertheless, more direct evidence is needed.

In the following experiment, we directly examined recognition memory for exemplars in rule-based and information-integration category learning. Recognition memory is commonly assumed to be a test of explicit memory processes (Gabrieli & Fleischman, 1995). If rule-based structures are learned explicitly and information-integration category structures are learned implicitly, as predicted by COVIS, then one would predict, if anything, better recognition memory performance for participants in the rule-based condition than for participants information-integration condition.

In contrast, our hypothesis is that that participants learn information-integration category structures explicitly using simple rules bolstered by memory for exceptions to those rules. This strategy would allow participants in the information-integration condition to score as highly as if they used the optimal diagonal decision bound, however it would also increase demands on recognition memory as participants would have to remember the exceptions. An alternative similar hypothesis is that participants in the information-integration condition may be using complicated rule strategies such as a conjunction rule. If this is the case, participants would still have to pay more attention to stimulus features information-integration than in the unidimensional condition as they would have to compare each stimulus to multiple boundaries. Whereas, for the unidimensional condition they only have to focus on one stimulus dimension. Either

way, we would expect to see better recognition memory in the information-integration condition than in the rule-based condition.

# Method

## Participants

Forty-two undergraduate psychology students were recruited from the Plymouth University participation pool and compensated with partial course credit.

## Stimuli and category structures

The stimuli were 36 grey squares that varied in brightness and size displayed on a white background. The stimuli seen by each participant depended on which category structure they learned.

Half the participants were randomly assigned to learn a unidimensional rule-based category structure and the other half to learn an information-integration category structure as illustrated in Figure 1. The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness (11 participants) or size of the square stimuli (10 participants)–and two information-integration category structures–where the optimum boundary had either a positive (10 participants) or negative gradient (11 participants). In addition, the stimuli were log-scaled so that all adjacent stimuli were approximately equally perceptually discriminable.

The abstract representation of the information-integration positive category structure is identical to that used by Spiering and Ashby (2008) with 6 stimuli added to bring the total number of stimuli up to 36. These stimuli were added to facilitate the random selection of a third of stimuli as "unseen" items for the recognition task. The remaining category structures are rotations ($\pi/4$, $\pi/2$, $3\pi/4$ rad) of this original structure around the origin and then translated so that 'center of gravity' of the points remained the same.

## Procedure

The experiment was split into three phases: category training, recognition test and finally, category test.

**Category training** In this phase, participants were trained on two thirds of the available stimuli. The training stimuli were selected randomly for each participant subject to several constraints: 1) that those stimuli selected were symmetrical around the category boundary and 2) that no adjacent stimuli of similar difficulty were removed. In total there were 360 training trials, split into 3 blocks of 120 trials. In each block, the 24 stimuli were shown 5 times in a random order. On each trial, the participants looked at the stimulus until they made a response using either the "Z" key for Category A or the "/" key for Category B. Participants were unable to respond until at least 500ms had passed. Then, either "Correct" in green or "Incorrect!" in red was displayed for 500ms. A blank white screen was displayed between each trial for 500ms. Throughout the experiment, the labels "Category A" and "Category B" were displayed on the bottom left and right of the screen respectively. If participants took longer than 5000ms to respond, no corrective feedback was given, instead the message "PLEASE RESPOND FASTER" was displayed for 500ms.

**Recognition test** In this phase participants judged whether each stimulus was "old" and appeared in the training phase, by pressing the "O" key, or was "new" and had not been shown in the training phase, by pressing the "N" key. The words "New" and "Old" were presented on the bottom left and right of the screen respectively. After this, participants judged the confidence they had in their old-new judgement on a Likert scale that varied from 1 (=guessed) to 5 (=certain) by pressing the corresponding number key. Each of the 36 stimuli were presented three times in a randomised order. No feedback was given.

**Category test** In this phase, participants were asked to judge the category membership of all 36 stimuli, not just those they had seen in the category training phase. The procedure was identical to that of the training phase, apart from there was no feedback. Each of the 36 stimuli were presented three times in a random order.

**Verbal report questionnaire** At the end of the experiment, participants were asked to complete a questionnaire that asked them to describe in detail the strategy that they used. This was to determine whether the participants could explicitly report the strategy they used and whether any participants used a rule-plus-exception strategy. The questionnaire asked them to "Imagine that another participant was asked to complete the experiment as you did. What instructions would you give them so that they could exactly copy your pattern of responding?"

The verbal reports were coded by CERE and AJW.

## Analysis

All data analyses were conducted in R (R Core Team, 2014). For every condition and phase of the experiment, all trials were removed for which the reaction times were outliers in that condition (i.e. outside 1.5 times the interquartile range above the upper quartile and below the lower quartile).

# Results

One participant was removed from the unidimensional condition because their accuracy score was consistently below chance (i.e. 50%), resulting in 20 and 21 participants in the unidimensional and information-integration category structure conditions respectively.

## Performance

**Category learning** We found a statistically significant difference in categorization accuracy at test, $F(1,39) = 13.51$, $p < .001$. Proportion correct was higher for the unidimensional category structure, $M_{UD} = 0.87$, $SD = 0.07$, than for

the information-integration category structure, $M_{II} = 0.78$, $SD = 0.11$.

**Recognition** To determine overall memory performance, we calculated $d'$ values for each participant.

We found that there was a statistically significant difference in $d'$ between the two category structure conditions, $t(39) = 2.04$, $p = .048$. Specifically, $d'$ was higher for the information-integration category structure, $d' = 0.01$, $SD = 0.02$, than for the rule-based category structure, $d' = 0.00$, $SD = 0.01$. Further, $d'$ was significantly greater than chance in the information-integration category structure condition, $t(20) = 2.98$, $p = .007$, but not in the unidimensional category structure condition, $t(19) = 0.67$, $p = .511$.

## Strategy analyses

The performance analyses above indicate a slight memory advantage for stimuli in the information-integration category structure condition compared to those in the unidimensional category structure condition. In this section, we investigate possible sources of this advantage.

**Model-based analysis** One possibility consistent with COVIS is that the category structure manipulation failed to result in a corresponding shift in category learning system. In other words, participants in the information-integration condition could have been using the sub-optimum, explicit system. If this were the case, then it would not be surprising that participants had explicit memory for category information. This is always a concern for experiments in the COVIS literature. The usual solution is to conduct a model-based analysis to determine which strategies participants are using to learn the structure. If the majority of participants in the information-integration condition are identified by the analysis as using the optimum diagonal strategy, then proponents of COVIS would conclude that those participants are using the implicit system.

In the model-based analysis typically used in the COVIS literature (Ashby & Gott, 1988), four types of model are fitted to the data from each participant. These model types are qualitatively different types of optimum decision boundaries that split the stimulus space into two, with "Category A" responses on one side and "Category B" responses on the other.

The *unidimensional* models assume that the stimuli are categorised on the basis of a single stimulus dimension. In this case, there are two possible unidimensional models: the stimuli can be split either on the basis of brightness or size. A unidimensional rule based on brightness would be "Place the light squares in Category A and the dark ones in Category B". This would be represented in stimulus space as a vertical or horizontal line. This model has two parameters: the value at which the boundary crosses the axis and perceptual noise.

The *conjunction* models assume that participants make a decision for each stimulus dimension and then combines them to determine category membership. A conjunction rule in this case might be "Place the light, small squares in Category A. Everything else is in Category B." This model can have up to three parameters: a decision criterion on each dimension and a noise parameter. Four versions of the conjunction model were included corresponding to each corner of the stimulus space.

The *general linear classifier* (GLC) models assume that the decision boundary between the categories can be described by a diagonal line between them. This is the optimum strategy for the information-integration condition. This model can have up to three parameters: the gradient, intercept and a noise parameter.

Two types of *random* models were also included: unbiased and biased. In the unbiased model, it is assumed that for every stimulus the participant is equally likely to pick either category. There are no parameters in this model. In the biased random model, it is assumed that for every stimulus the participant in likely to ascribe it to Category A with a certain probability (i.e. 30%). This model has one parameter, the proportion of Category A responses, and is a more general version of the unbiased random model, for which the parameter is equal to 50%.

The data from each participant was fitted to each of these models. The degree of fit was measured by the Bayesian Information Criterion (Schwarz, 1978). The results from this analysis, which was performed using the grt package in the R environment Matsuki (2014), are reported in Table 1.

Here we can see that our data meets the criterion commonly used in the COVIS literature: the majority of participants in the information-integration condition have been found to be using the optimum GLC (or diagonal) strategy. Researchers in the COVIS framework would normally conclude from this that people in the information-integration condition were using the optimum implicit system.

Table 1: Strategies identified in each condition according to the model-based analysis.

| | Strategies | | | |
|---|---|---|---|---|
| Condition | UD | GLC | CJ | RND |
| UD | 14 | 2 | 3 | - |
| II | 6 | 10 | 4 | 1 |

Category structures: UD=Unidimensional, II=Information-integration. Models: UDX=Unidimensional based on brightness, UDY=Unidimensional based on size, GLC=General linear classifier, CJ=Conjunction RND=Random (both types).

**Verbal reports** The model-based analysis indicates that our data meet the minimum requirements for a COVIS study: there was an obvious effect of strategy type depending on which category structure participants learned. However, as mentioned in the introduction, the results of this analysis have been found to depend on the models included in the analysis

Table 2: Strategies identified in the verbal report questionnaire

| | Strategies | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | UD | seqUD | CJ | CJ2 | RuleX | Implicit | Other | None |
| UD | 15 | 1 | 1 | 0 | 0 | 1 | 0 | 2 |
| II | 4 | 4 | 3 | 5 | 1 | 0 | 1 | 4 |

Category structures: UD=Unidimensional, II=Information-integration. Strategies: UD=Unidimensional, seqUD=Sequential unidimensional, CJ=Conjunction, CJ2=Double conjunction, RuleX=Rule-plus-exception.

(Donkin et al., 2015). Additionally, our previous research found that participants identified as diagonal classifiers could also report using complex rule-based strategies (Edmunds et al., 2015). In this section we examine the strategies that participants report using to see if participants in the information-integration condition report using complex rule-based strategies.

Participants reported several different types of strategy. A report was classified as a *unidimensional* strategy if it described a rule based only of the stimulus dimensions such as "the dark squares are Category A and the light ones are Category B."

A report was classified as a *sequential unidimensional* strategy if the participants first categorised the stimuli at the extreme ends of one stimulus dimension and then defined a second unidimensional rule, on the other stimulus dimension, for the stimuli in the middle of the dimension. For example, a participant might say: "The very small stimuli were in Category A, and the very large in Category B. For the middle sized stimuli, the light ones were in Category A and the dark in Category B."

A report was classified as a *conjunction* strategy if the participant described an AND rule on the basis of two stimulus dimensions such as "Stimuli that are both small and dark are in Category A; else Category B."

A report was defined as a *double conjunction* strategy if the participant described two opposing corners of the stimulus space, but failed to define the other areas of the space. For example: "Large and dark patterns go into B. Small and light colours into A." As can be seen in this example, the participant fails to describe what category small dark stimuli would be in.

A report was classified as a *rule-plus-exception* strategy if the participant reported a simple rule with some exceptions. For example, "Light stimuli were usually Category A and dark stimuli Category B. However, one light medium sized stimulus was in Category B."

A report was classified as an *implicit* strategy if the participant recommended "not thinking too much" or to "rely on instinct" or similar phrases.

The inter-rater reliability for whether or not a participant reported a strategy was perfect. Similarly, both coders agreed perfectly on the strategies participants used.

As we can see in Table 2, no participants reported using an implicit strategy in the information-integration condition. This replicates our finding that the model-based analysis does not correspond well to the strategies participants report (Edmunds et al., 2015). Furthermore, only one participant in the information-integration condition used a rule-plus-exception strategy. This indicates that the advantage in memory may not be due to the use of a particular strategy, but because of the need in complex strategies to attend closely to the stimuli in order to categorise them by comparing to multiple decision bounds.

## Discussion

### Summary

A key dual-system model of category learning, COVIS, predicts that categorization is mediated by two competing learning systems: one explicit and one implicit. These two systems are hypothesised to optimally learn two different types of category structure. The explicit verbal system optimally learns rule-based category structures, whereas the implicit system optimally learns information-integration category structures. A key feature of this model is that category knowledge learned using the implicit system is unavailable to consciousness (Smith et al., 2015). In contrast, behavioral and neuroscientific work from out lab indicates that participants learning information-integration categories are aware of category knowledge and may be using explicit memory to facilitate categorization (Carpenter et al., in press; Edmunds et al., 2015). This experiment aimed to directly test these possibilities by comparing participants' performance on an old-new recognition task after learning either a unidimensional rule-based category structure, or an information-integration one.

Contrary to the predictions of COVIS, we found superior memory for exemplars after learning an information-integration category structure compared to a rule-based one. This indicates that participants may learn information-integration category structure using complex rule-based strategies rather than implicitly. This would result in superior exemplar memory for items when learning a information-integration category compared to a unidimensional structure as participants would have to attend more closely to the stimuli's features in order to compare them to multiple decision boundaries. Previous conclusions in the COVIS literature

concerning the presence of implicit-like category learning may be due to an over-reliance on the assumption that a limited model-based analysis can provide evidence for implicit responding (Donkin et al., 2015).

One apparent limitation of the current study is that response accuracy in the category test phase is lower in the information-integration condition than in the rule-based condition. In an ideal comparison between rule-based and information-integration learning, the conditions would be matched for error rate. However, it seems likely that improving overall performance on the information-integration could only improve memory for the exemplars as this would involve using a more refined rule-based strategy. Thus, it seems unlikely that such a change would qualitatively alter our conclusions.

Another potential limitation is that recognition performance is poor in both conditions of the current experiment. This may be due to the stimuli being perceptually very similar to one another. Further work might increase discriminability by adding additional features that were not predictive to category membership. We are currently investigating this possibility.

In conclusion, this experiment finds preliminary evidence that participants learning the information-integration category structure do so explicitly. This conclusion is in contrast to the prediction of the COVIS model, which assumes that the information-integration structure is learned implicitly.

## References


Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 1–13). New York: Cambridge University Press.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. N. (in press). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 933–948.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures . *The Quarterly Journal of Experimental Psychology*, *68*(2), 1203–1222.

Gabrieli, J. D. E., & Fleischman, D. (1995). Double dissociation between memory systems underlying explicit and implicit memory in the human brain. *Psychological Science*, *6*(2), 76–82.

Matsuki, K. (2014). grt: General recognition theory [Computer software manual].

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*(8), 1326–1336.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). *Systems of Category Learning. Fact or Fantasy?* (Vol. 54). Elsevier Inc.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, a. D., Gitelman, D. R., Parrish, T. B., ... Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568.

Price, A., Filoteo, J. V., & Maddox, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*(5), 1213–1226.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*(7), 2476–2490.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information integration, but not rule-based category learning. *Psychological Science*, *19*(11), 1169–1177.

Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*(7), 1747–1758.

Zaki, S. R., & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: evidence for multiple systems or task complexity? *Memory & Cognition*, *42*(3), 508–24.