ORIGINAL ARTICLE

# Category label and response location shifts in category learning

**W. Todd Maddox · Brian D. Glass · Jeffrey B. O'Brien ·
J. Vincent Filoteo · F. Gregory Ashby**

**Abstract** The category shift literature suggests that rule-based classification, an important form of explicit learning, is mediated by two separate learned associations: a stimulus-to-label association that associates stimuli and category labels, and a label-to-response association that associates category labels and responses. Three experiments investigate whether information–integration classification, an important form of implicit learning, is also mediated by two separate learned associations. Participants were trained on a rule-based or an information–integration categorization task and then the association between stimulus and category label, or between category label and response location was altered. For rule-based categories, and in line with previous research, breaking the association between stimulus and category label caused more interference than breaking the association between category label and response location. However, no differences in recovery rate emerged. For information–integration categories, breaking the association between stimulus and category label caused more interference and led to greater recovery than breaking the association between category label and response location. These results provide evidence that information–integration category learning is mediated by separate stimulus-to-label and label-to-response associations. Implications for the neurobiological basis of these two learned associations are discussed.

W. T. Maddox (✉) · B. D. Glass
Department of Psychology, Institute for Neuroscience,
University of Texas, 1 University Station,
A8000, Austin, TX 78712, USA
e-mail: maddox@psy.utexas.edu

J. B. O'Brien · F. G. Ashby
University of California, Santa Barbara, USA

J. V. Filoteo
VA San Diego Healthcare System,
University of California, San Diego, USA

## Introduction

An important topic of psychological research is to examine the cognitive processes that allow people to be flexible in their behavior and to adapt to novel or changing situations. One popular method for examining these processes is to train people on a task and then to examine the performance costs and recovery rates associated with various "shifts" in the nature of the problem. A paradigm that has been used extensively to study this problem is *rule-based classification*. In rule-based classification the rule that maximizes accuracy (i.e., the optimal strategy) is easy to describe verbally and can be learned via an explicit reasoning process (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Bruner, Goodnow, & Austin, 1956; Estes, 1994; Smith & Medin, 1981). Rule-based tasks can be contrasted with information–integration, family resemblance, and other types of ill-defined classification tasks for which the optimal strategy involves some implicit integration of information across stimulus dimensions (Ashby et al., 1998; Milton, Longmore, & Wills, 2008; Neisser, 1967; Smith & Medin, 1981; Wills, Noury, Moberly, & Newport, 2006).

In the application most often used in the rule-based shift literature, one stimulus dimension is relevant, and the participant's task is to discover the relevant dimension and then to map the different dimensional values to the relevant categories. We refer to these as one-dimensional rule-based tasks. Once the one-dimensional rule is learned, the category labels might be reversed (a reversal shift), new values along the same relevant dimension might be substituted (an

intra-dimensional shift), or a previously irrelevant dimension might become relevant (an extra-dimensional shift) (Buss & Buss, 1956; Downes et al., 1989; Goldstone & Steyvers, 2001; Kendler & Kendler, 1962; Kruschke, 1996; Robbins, 2007; Wills et al., 2006; Wolff, 1967).

One critical finding that emerged from early work on rule-based shifts is that reversal shifts are easier to learn than extra-dimensional shifts (e.g., Buss, 1953). Kendler and Kendler (1962, 1968) argued that these data support the existence of a category representation that mediates between the stimulus and the category label. Reversal shifts are easy to learn because the association between the stimuli and this intermediate category representation can be modified quickly. While these data are congruent with a two-association (stimulus-to-category representation and category representation-to-category label) model, they are also congruent with a single association attentional model that assumes that training increases attention to the relevant dimensions and decreases attention to the irrelevant dimensions (Sutherland & Mackintosh, 1971).

Some of the clearest evidence in support of separate associations, over a purely attentional explanation comes from Sanders (1971) and Wills et al. (2006) who compared performance across full and partial reversal conditions. Sanders study utilized rule-based categories, whereas Wills et al. used family resemblance categories. The full reversal condition was identical to the standard reversal shift paradigm. However, in the partial reversal condition, the stimulus-to-category label associations reversed for some stimuli, but not for others. As expected from a two-association model, the partial reversal was more difficult to learn than the full reversal. Taken together, these studies suggest that the learning of stimulus-to-category label associations actually involves learning a stimulus-to-category representation association and a category representation-to-category label representation.

To further explore the nature and flexibility of rule-based classification and to build upon the seminal work of Sanders (1971) and Wills et al. (2006), Kruschke (1996) examined rule-based shift learning when the initial classification required an exclusive-or (XOR) rule on two dimensions with a third irrelevant dimension. A number of shifts were examined, including reversals, shifts to a one-dimensional rule on a previously relevant dimension, shifts to a one-dimensional rule on a previously irrelevant dimension, and shifts to a new XOR rule on one previously relevant and one previously irrelevant dimension. Kruschke reported that relearning was fastest in the reversal condition, followed by the shift to a one-dimensional rule on a relevant dimension, a shift to a one-dimensional rule on a previously irrelevant dimension, and a shift to a new XOR rule.

Based on these results, Kruschke proposed that dimensions relevant during training have heightened attention at

transfer, making it easier to learn new categories based on the same relevant dimensions. Kruschke also proposed that training strengthens category representations, which makes transfer easier when only the category label changes (e.g., as in a reversal) compared to transfer that requires learning new category representations. This latter principle is related to the notion of "mediating responses" proposed by (Kendler & Kendler, 1962) to account for the ease of reversal shift learning (i.e., the notion that a category representation exists that mediates between the stimulus and category label).

To account for these data, Kruschke (1996) proposed a four-layer connectionist model (AMBRY) that included an input, exemplar, category representation, and category response layer. In his original ALCOVE model, the category representation and category response layers were combined (Kruschke, 1992). By separating the category representation/response layer into a category representation layer that receives input from the exemplar layer, and a fourth response layer that receives input from the category representation layer the model can account for the differential effects of shift manipulations that affect the exemplar-to-category label connections by changing the set of stimuli associated with the category representation (e.g., shifts to one-dimensional rules or a new XOR rule) from shifts that affect the category label-to-response connections (e.g., reversals). For ease of exposition, we refer to the stimulus (or exemplar) to category label connections as the *Stimulus-to-Label* associations, and the category label-to-response connections as the *Label-to-Response* associations.

The focus of the current study is on determining whether the stimulus-to-label and label-to-response associations thought to characterize rule-based classification, are also associated with implicit information–integration classification. Experiments 1 and 2 examine conditions in which each category label is associated with a single stimulus cluster. This approach effectively sidesteps the issue of whether the stimulus-to-label association can be decomposed into a stimulus-to-category representation association and category representation-to-category label association, ala Sanders (1971) and Wills et al. (2006). We took this approach to focus instead on dissociating the stimulus-to-label and label-to-response associations in information–integration category learning. Experiment 3 examines a situation similar in spirit to that taken by Sanders (1971) and Wills et al. (2006). In the General discussion, we address the possibility that classification might be characterized by at least three sets of learned associations: stimulus-to-category representation, category-representation-to-category label, and category label-to-category response.

Rule-based classification is an important domain within which to study the effects of shifts on performance because rule-based classification is ubiquitous in daily life, and

because it recruits a general form of explicit learning. Even so, other types of classification that seem to depend on implicit, rather than explicit learning are common in the real-world. One that has been studied extensively in the categorization literature, but to date has not been examined using shift procedures (however, see Wills et al., 2006), is the information–integration classification task (Ashby et al., 1998; Maddox & Ashby, 2004).

*Information–integration* category structures are those in which optimal accuracy requires integrating information from two or more stimulus dimensions (usually expressed in different physical units). An example of information–integration categories composed of circular sine-wave gratings is shown in Fig. 1. The optimal strategy (denoted by the solid diagonal line) could be verbalized as "respond A when the orientation is greater than the bar width; otherwise respond B," but this is impossible to comprehend because it compares dimensions expressed in incommensurable units. It is impossible to directly compare an orientation, measured in degrees with a bar width, measured in inches, and to determine when one is "greater than" the other.

We expect that the effects of category shifts on information–integration classification might be fundamentally different than on rule-based classification for at least two reasons. First, Ashby, Ell, and Waldron (2003) already examined the effects of switching the response keys after an initial training period in rule-based and information–integration classification. Note that such a switch disrupts the label-to-response mapping, but not the stimulus-to-label mapping. Switching the keys caused no interference for rule-based categories, but had a large effect for information–integration strategies. Thus, at least in this one study, a manipulation that affected the category label-to-response association differentially affected rule-based and information–integration categories. The goal of the current study is to replicate and extend this effect to different types of rule-based and information–integration categories, and to extend
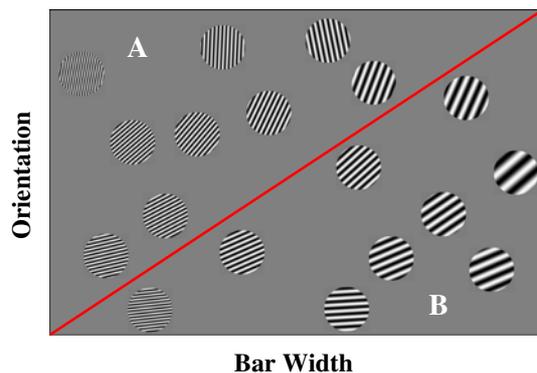


**Fig. 1** An example of category structures that might be used in an information–integration category learning task. The optimal decision bound is indicated by the *diagonal line*

this to manipulations that affect the stimulus-to-category label associations.

Second, there is much evidence that information–integration classification recruits different cognitive and neural systems than rule-based classification. Briefly, rule-based classification appears to recruit explicit executive processes that are mediated primarily within frontal cortex, whereas information–integration classification recruits implicit procedural-learning systems that are largely mediated by the striatum, a brain region known to be directly involved in motor and response learning (Filoteo et al., 2005; Nomura & Reber, 2008; Nomura et al., 2007; Poldrack & Foerde, 2008; Poldrack et al., 2001; Seger, 2008; Seger and Cincotta 2005, 2006). It is likely that the effects of category shifts will differ as a function of the cognitive and neural systems that sub-serve rule-based and information–integration categorization.

In summary, by applying shift manipulations to information–integration classification we extend our understanding of the effects of category shifts to non-verbal forms of classification in particular, and to more implicit forms of learning in general. In addition, this approach allows us to determine whether the two-association (stimulus-to-label and label-to-response) model that characterizes rule-based classification also characterizes information–integration classification.

### Overview of present studies

The overriding goal of this research was to confirm the existence of separate stimulus-to-label and label-to-response associations, hypothesized in previous work (Goldstone & Steyvers, 2001; Kendler & Kendler, 1962; Kruschke, 1996), in conjunctive, rule-based learning and to determine whether these same set of learned associations is present in information–integration category learning. The results from three experiments are reported. Experiment 1 examined two types of category shifts in rule-based classification learning and can be thought of as an extension of Kruschke (1996). Experiment 2 was a direct replication of Experiment 1 except that rule-based categories were replaced with information–integration categories. Importantly, the rule-based and information–integration category structures used in Experiments 1 and 2 are related by a simple linear rotation in the stimulus space. Thus, the two-category structures are equivalent on a number of important properties including within and between category coherence, optimal accuracy, etc. Experiment 3 examined a different information–integration category structure to test the generalizability of the Experiment 2 results.

During pre-change training in Experiment 1, participants learned four-categories that could be classified with a conjunctive rule. The stimuli comprising each category were

lines that varied in length and orientation. A scatterplot of the stimuli is displayed in the top of Fig. 2, along with the optimal decision bounds. The open and filled squares and triangles denote the stimuli from four separate stimulus clusters. Note that the task is conjunctive and rule-based because the rule is easily verbalized as: give one response to short, shallow angle lines, another to short, steep angle lines, a third to long, shallow angle lines and a fourth to long, steep angle lines.

The assignment of stimulus clusters (from Fig. 2) to category labels and response locations in the various conditions is outlined in Fig. 3. In Fig. 3, the "Stimulus Cluster" column denoted the symbol associated with each of the four clusters of stimuli. The "Category Label" column denotes the labels, A–D, and the "Response Location" column
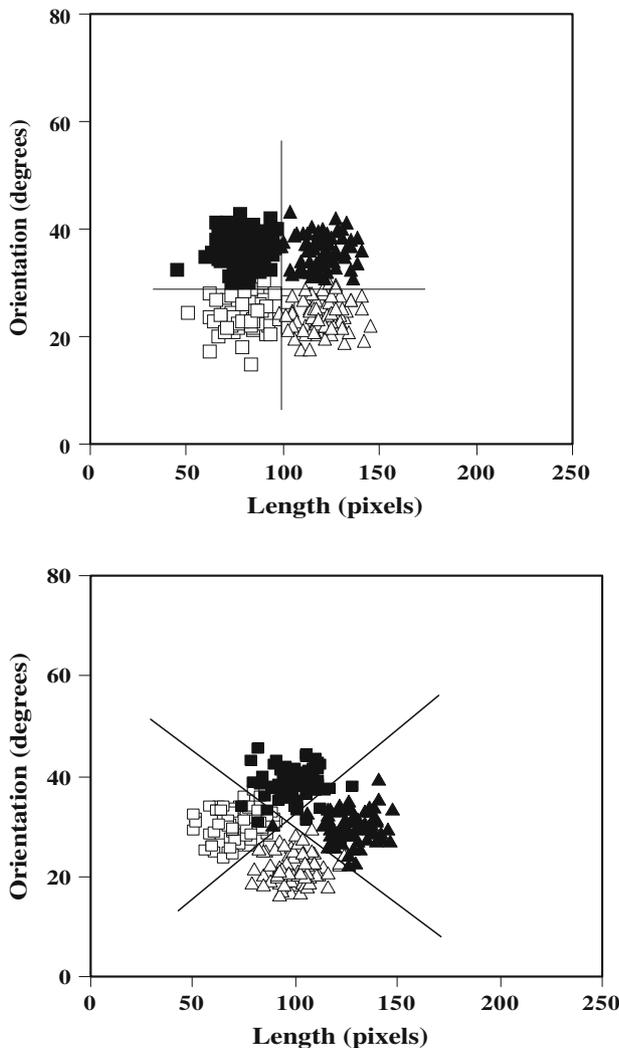


**Fig. 2** Categorization conditions used for Experiments 1 and 2. The stimuli used in all conditions from Experiment 1 are displayed in the *top panel*. The stimuli used in all conditions from Experiment 2 are displayed in the *bottom panel*. *Solid lines* denote the optimal decision bounds, and the *open squares*, *filled squares*, *open triangles*, and *filled triangles* denote the stimuli



**Fig. 3** Experimental conditions from Experiments 1 and 2. The stimulus-to-category label assignments are denoted by the *lines* that connect the stimulus cluster symbols (from Fig. 2) with a category labels. Notice that these assignments are the same in the control and response location conditions, but differ in the category label condition. The category label-to-response location assignments are denoted by the *lines* that connect the category labels with response locations. Notice that these assignments are the same in the control and category label conditions, but differ in the response location condition

denotes the buttons on the computer keyboard that are used throughout the experiment ("Z", "W", "/" and "P"). The assignment of stimulus clusters-to-category labels is denoted by the lines connecting each stimulus cluster to each category label. The assignment of category labels-to-response locations is denoted by the lines connecting each category label to each response location.

During pre-change training in the three experimental conditions, Control, Category Label and Response Location, a fixed stimulus-to-category label and category label-to-response location mapping was used (see Fig. 3 for details). During post-change training in the Control condition the same stimulus-to-category label and category label-to-response location mappings were used. Two versions of

the Category Label condition were examined. Only version A is depicted in Fig. 3. In the Category Label(A) condition, the association between stimulus clusters and category labels was changed during post-change training so that stimuli originally assigned to categories A, B, C, and D were now assigned to categories B, A, D, and C, respectively (see Fig. 3). In the Category Label(B) condition, the association between stimulus clusters and category labels was changed during post-change training so that stimuli originally assigned to categories A, B, C, and D were now assigned to categories C, D, A, and B, respectively (see Fig. 3). Note that the lines connecting the stimulus cluster to the category label are different in the Category Label conditions than in the Control condition. These instantiations of the category label manipulation represents a four-category analog of a *reversal shift*. Importantly, although the stimulus-to-category label associations were modified in the Category Label condition, the response locations on the computer keyboard associated with category label A (the Z key), B (the W key), C (the / key), and D (the P key) remained unchanged. This lack of change in the category label-to-response location assignments can be seen in Fig. 3 by noting that the lines connecting the category label to the response location do *not* differ across the Category Label and Control conditions.

Notice that the Category Label(A) manipulation switches the stimulus-to-category label associations along the orientation dimension. The Category Label(B) manipulation switches the stimulus-to-category label associations along the length dimension. By examining both conditions, we effectively counterbalanced this factor. We took the same counterbalancing approach with respect to the response location manipulation (see below). To anticipate, the performance profiles for the A and B versions of the experimental manipulation did not differ statistically and thus we collapsed across this factor when describing the major findings.

Two versions of the Response Location condition were also examined. In the Response Location(A) condition, the category label-to-response location associations changed during post-change training so that the buttons "Z", "W", "/", and "P" that were originally associated with category labels A–D, respectively, changed in such a way that buttons "Z", "W", "/", and "P" were now associated with category labels B, A, D, and C, respectively. In the Response Location(B) condition, the category label-to-response location associations changed during post-change training so that the buttons "Z", "W", "/", and "P" were now associated with category labels C, D, A, and B, respectively. Note that the lines connecting each category label with each response location differ across the Response Location and Control conditions. Importantly, although the category label-to-response location assignments were modified in

the Response Location condition, the stimulus-to-category label associations remained unchanged. This lack of change in the stimulus-to-category label assignments can be seen in Fig. 3 by noting that the lines connecting the stimulus cluster to the category label do *not* differ across the Response Location and Control conditions.

Several comments are in order regarding the Category Label and Response Location conditions. First, note that in both conditions, the stimulus-to-response associations that were developed during pre-change training were broken during post-change training so that each stimulus cluster required a different button press post- versus pre-change. Thus, a model that assumes a single (stimulus-to-response) association would predict no performance difference across the Category Label and Response Location conditions, because both conditions break the learned stimulus-to-response association.

Second, it is important to note also that the stimuli that cluster together into a single group (or category) remained unchanged in all conditions. Only the category label associated with each group of stimuli (Category Label condition), or the response button associated with that label changed (Response Location condition). Thus, these manipulations differ from the intra- or extra-dimensional shifts typically used in the literature because those shifts actually change the structure of the categories and the nature of the optimal decision bound. These manipulations also differ from the partial shift manipulations used by Sanders (1971) and Wills et al. (2006) because all stimuli originally trained to a particular category label, either remain the same (as in the Response Location condition) or *all* change (as in the Category Label condition). In this way, our two manipulations are identical with respect to the input and output states of the system in the sense that the optimal decision bounds remain the same (although the stimulus-response mappings change), and the stimulus-response mappings for all stimuli in each stimulus cluster change in the same way. Comparing performance across the Category Label and Response Location conditions will provide a powerful empirical test to determine whether processes associated with each of these two mappings differs under these fundamental but fairly subtle manipulations.

Whereas the single, stimulus-response association model predicts no difference in the performance profile across the Category Label and Response Location conditions, the two-association model (stimulus-to-label association and label-to-response association) could predict a larger performance cost for the category label manipulation relative to the response location manipulation or vice versa. Although the specific category label and response location manipulations outlined in Fig. 3 have not been examined in the literature, the most likely prediction to follow from the extant literature (Kruschke, 1996; Wills et al., 2006) is that the cost

associated with the category label manipulation will be larger than that for the response location manipulation.

We turn now to Experiment 1 that examines the effects of category label and response location manipulations on four-category conjunctive rule-based classification learning.

## Experiment 1

In Experiment 1, participants learned the four-category conjunctive rule-based task described in the top of Fig. 2. Participants in the Control, Category Label and Response Location conditions completed three 100-trial pre-change training blocks, followed by three 100-trial post-change transfer blocks. Experiment 1 extends Kruschke (1996) to category shift conditions that break the stimulus-to-category label associations (Category Label condition) or category-label-to-response location associations (Response Location condition).

Methods

### Participants

One hundred and eleven participants completed the study and received course credit for their participation. All participants had normal or corrected to normal vision. Each participant served in one condition. To ensure that only participants who performed well above chance were included in the post-change performance analyses, a learning criterion of 40% correct (25% is chance) during the final pre-change block was applied. All but ten participants met the performance criterion (Control $N = 20$; Category Label $N = 40$; Response Location $N = 41$), with approximately equal numbers of participants completing the A and B versions of the Category Label and Response Location manipula4ions.

### Stimuli and stimulus generation

The stimuli are displayed in Fig. 2, and were generated by drawing 25 random samples from each of four bivariate normal distributions along the two stimulus dimensions with means along the $x$ dimension of 80, 80, 120, and 120 and along the $y$ dimension of 80, 120, 80, and 120 for categories A–D, respectively. The variance along the $x$ and $y$ dimension was 100 and the covariance was 0 for all categories. The random samples were linearly transformed so that the sample means and variances equaled the population means and variances. Each random sample $(x, y)$ was converted to a stimulus by deriving the length (in pixels) as $l = x$, and orientation (in degrees counterclockwise from

horizontal) as $o = y - 30$. These scaling factors were chosen to roughly equate the salience of each dimension. Optimal accuracy was 95%. The 100 stimuli were randomized separately for each participant in each block during the pre-change trials for all conditions and for the post-change Control and Response Location condition trials. For the post-change Category Label condition trials, the same stimuli were used, but the category labels were changed in the way depicted in Fig. 3.

### Procedure

Participants were randomly assigned to one of the five experimental conditions: Control, Category Label(A or B) or Response Location(A or B). Each condition consisted of 3, 100-trial pre-change training blocks, followed by 3, 100-trial post-change transfer blocks with a participant-controlled rest period between each block.

During the Pre-Change blocks of all conditions, participants were told that they were to categorize lines on the basis of their length and orientation, that there were four equally likely categories, and that high levels of accuracy could be achieved. At the start of each trial, a fixation point was displayed for 1 s and then the stimulus appeared. The stimulus remained on the screen until the participant generated a response by pressing the "Z" key for category A, the "W" key for category B, the "/" key for category C, or the "P" key for category D. None of these four keys were given special labels. Rather the written instructions informed participants of the category label to button mappings, and if any button other than one of these four was pressed, an "invalid key" message was displayed. Following the response, the word "correct" was presented if their response was correct or the word "incorrect" was presented if their response was incorrect, along with the correct category label. Once feedback was given, the next trial was initiated.

The Post-Change Control participants completed three additional 100-trial transfer blocks. Following pre-change training, post-Change Category Label and post-Change Response Location participants were instructed that the stimuli associated with each category label had changed, the assignment of categories to buttons had changed, or both, and that they would have to learn the task from trial-by-trial feedback. The then completed three additional 100-trial transfer blocks.

Results

We first focus on standard statistical analyses of the accuracy data and then we consider model-based analyses. Three accuracy measures were examined. First, we examined *pre-change performance* to verify that no differences emerged across the three conditions, with an emphasis on

performance during the final pre-change block. Second, we examined the *cost* associated with transfer by subtracting accuracy in the first post-change block from accuracy in the final pre-change block. For more fine-grained detail, we also examined the cost based on the first 50 post-change trials, as opposed to the full 100-trials in the first post-change block. The larger the cost the greater the immediate impact on performance of the Category Label or Response Location switch. Third, for the Category Label and Response Location conditions, we examined *recovery* by subtracting accuracy in the first post-change transfer block of 100-trials from accuracy in the final post-change transfer block of 100-trials. The larger this value the greater the performance recovery during transfer.

*Accuracy analyses*

*Pre-change performance*   The learning curves for all three conditions across the three pre- and three post-change blocks are shown in Fig. 4a. A three condition × three pre-change block ANOVA was conducted to determine whether there were any pre-change performance differences. The block effect was significant $F_{(2, 196)} = 97.94$, $p < .001$, $\eta^2 = .500$], but the condition effect [$F_{(2, 98)} = .094$, ns, $\eta^2 = .002$] and the interaction [$F_{(4, 196)} = .28$, ns, $\eta^2 = .006$] were both non-significant. Most importantly, there were no differences across conditions in the final pre-

change block [$F_{(2, 98)} = .019$, ns, $\eta^2 = .000$]. Thus, pre-change training performance was equated across conditions.

*Cost*   The performance costs are displayed in Fig. 4b. The cost was significantly larger than zero in the Category Label condition [$t_{(39)} = 4.65$, $p < .001$], but did not differ from zero in the Response Location [$t_{(40)} = 1.73$, $p = .091$] or Control conditions [$t_{(19)} = .98$, $p = .339$]. The main effect of condition on the performance cost was significant [$F_{(2, 98)} = 4.39$, $p < .05$, $\eta^2 = .082$]. In addition, the cost was significantly larger in the Category Label condition than in the Response Location condition [$t_{(79)} = 2.69$, $p < .01$] and than in the Control condition [$t_{(58)} = 2.06$, $p < .05$]. The costs did not differ significantly across the Control and Response Location conditions ($t < 1.0$).

When we focused on the first 50 post-change trials in determining the cost, we observed performance costs of .010, .084, and .049 in the Control, Category Label and Response Location conditions, respectively. In this case the cost was significantly larger than zero in both the Category Label [$t_{(39)} = 5.52$, $p < .001$] and Response Location conditions [$t_{(40)} = 4.29$, $p < .001$], but not in the Control condition ($t < 1.0$). The main effect of condition on the performance cost was significant [$F_{(2, 98)} = 5.43$, $p < .01$, $\eta^2 = .100$], but the costs did not differ significantly across the Category Label and Response Location conditions



Fig. 4 **a** Proportion correct (averaged across participants) from Experiment 1. **b** Cost determined by subtracting post-change block 1 performance from pre-change block 3 performance. **c** Recovery determined by subtracting post-change block 1 performance from post-change block 3 performance. Standard *error bars* included
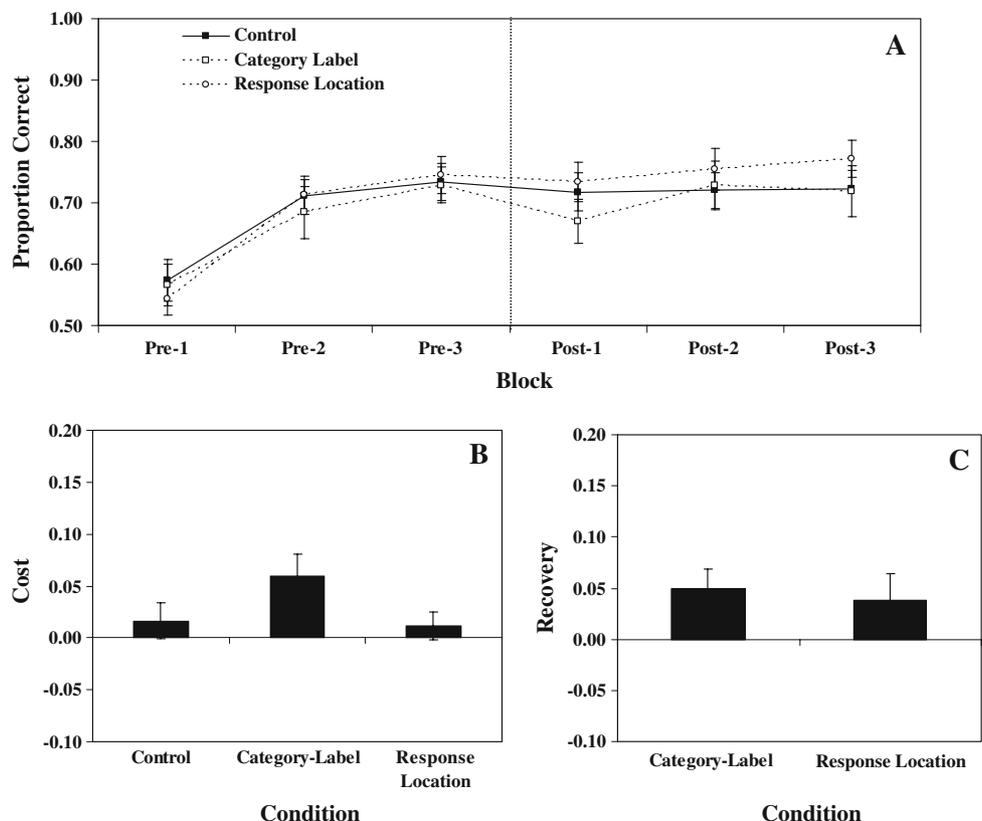
$[t(79) = 1.87, p = .065]$. The cost was larger in the Category Label condition than in the Control condition $[t(58) = 2.99, p < .01]$, but was not significantly larger in the Response Location condition than in the Control condition $[t(59) = 1.92, p = .06]$. Despite the lack of significance between the Category Label and Response Location costs based on the first 50 post-change trials, there is a clear trend toward a larger cost in the Category Label condition (.084) relative to the Response Location condition (.049), and toward a larger cost in the Response Location condition relative to the Control condition (.010).

*Recovery* The recovery data for the Category Label and Response Location conditions are displayed in Fig. 4c. Recovery was significant in the Category Label $[t(39) = 2.59, p < .05]$, and Response Location conditions $[t(40) = 3.91, p < .001]$. Even so, the difference in the magnitude of recovery across the two conditions was not statistically significant $[t(79) = .263, ns]$.

The most important finding is that the cost was larger in the Category Label condition than in the Response Location condition regardless of whether we focused on the first 50 post-change trials or the first 100 post-change trials. This difference did reach statistical significance based on the full 100-trial analyses, and revealed a clear trend based on the first 50-trial analyses. Taken together, these analyses suggest that manipulations that break the learned association between stimuli and category labels have a larger adverse effect on performance then manipulations that break the learned association between category labels and responses. This supports a two-association model of rule-based classification as suggested by the data from several researchers (e.g., Kendler & Kendler, 1962; Kruschke, 1996; Wolff, 1967). In addition, the cost was significantly larger than zero in the category label condition for both the 50- and 100-trials analyses. This suggests that a four-category variant of a reversal shift does lead to a performance cost. Finally, the effect of the response location manipulation was shorter lived, leading to a significant cost when we focused on the first 50 post-change trials, but not when we focused on the first 100 trials. This suggests that recover begins more quickly following a response location manipulation than following a category label manipulation. Ashby et al. (2003) found no performance cost during the first 50 post-change trials in a "button-switch" analog of our response location manipulation using a two-category one-dimensional rule-based task. Although further work is needed, the current data suggest that a button switch cost does emerge when the rule is conjunctive, and four-categories are relevant (see Maddox, Lauritzen, & Ing, 2007 for a related finding).

The accuracy-based analyses support for the claim that two associations characterize rule-based classification (e.g.,

Kendler & Kendler, 1962; Kruschke, 1996; Wolff, 1967), and extend this to a conjunctive rule-based tasks. Even so, it is important to determine whether the manipulations had any effect on the strategy that participants used to learn the categories. Our previous work with button switch manipulations suggests that participants will likely use rule-based decision strategies pre- and post-change in this task because the optimal strategy is rule-based (Ashby et al., 2003; Maddox & Ashby, 2004; Maddox et al., 2007). To address this issue decision bound models were fit on a block by block basis separately to the data from each participant (Maddox & Ashby, 1993). As expected, and in line with previous work (Ashby et al., 2003; Maddox et al., 2007), we found an overwhelming majority of the data sets were best fit by rule-based models. In fact, across all three conditions and all six blocks of trials, in no case did the percentage of data sets best fit by a rule-based model fall below 92%.

## Discussion

We examined the performance cost and recovery rate associated with experimental manipulations that broke the learned stimulus-to-category label association (Category Label condition) or the learned category label-to-response location association (Response Location condition). Most importantly, the magnitude of the cost was larger (significant based on the first 100 post-change trials and showing a strong trend based on the first 50 post-change trials) when the stimulus-to-category label association was broken than when the category label-to-response location association was broken. Despite the larger cost in the former case, the recovery rates did not differ statistically across conditions. In addition, participants were found to use rule-based strategies throughout the pre-change and the post-change performance blocks. These data lend support to proposals that there are at least two separate associations that are involved in rule-based classification learning. We turn now to an experiment that examines the effects of identical Category Label and Response Location manipulations on information–integration classification.

## Experiment 2

Experiment 2 replicates and extends the Category Label and Response Location manipulations used in Experiment 1 to a four-category information–integration task. Importantly, the information–integration categories were derived from a 45° rotation of the rule-based categories used in Experiment 1 (see Fig. 2). Thus, the categories are structurally equivalent across the two studies, and any performance differences must be attributed to the qualitative difference in the nature of the categorization strategies—that is, rule-

**Fig. 5** **a** Proportion correct (averaged across participants) from Experiment 2. **b** Cost determined by subtracting post-change block 1 performance from pre-change block 3 performance. **c** Recovery determined by subtracting post-change block 1 performance from post-change block 3 performance. Standard *error bars* included
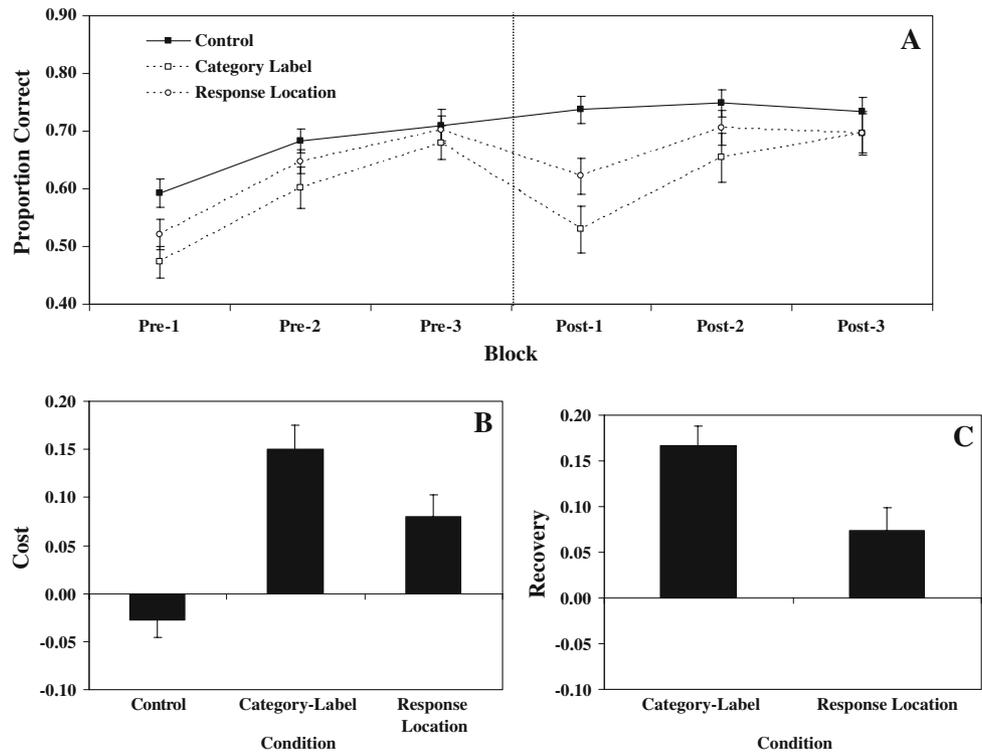


based versus information–integration. The results from Experiment 2 will allow us to determine whether the two associations that characterize rule-based classification learning, also characterize information–integration classification learning, and will allow us to compare and contrast the processing characteristics of the two associations across rule-based and information–integration tasks.

## Methods

### Participants

One hundred and five participants completed the study and received course credit for their participation. All participants had normal or corrected to normal vision. Each participant served in one condition. Of the 105 participants, 97 met the learning criterion of 40% in the final pre-change block (Control $N = 19$; Category Label $N = 37$; Response Location $N = 41$), with approximately equal numbers of participants completing the A and B versions of the Category Label and Response Location manipulations.

### Stimuli and stimulus generation

The category structures are displayed in Fig. 2. Stimuli were generated from the Experiment 1 stimuli via a 45° rotation. This yielded means along the $x$ dimension of 72, 100, 100, and 128 and along the $y$ dimension of 100, 128, 72, and 100 for categories A–D, respectively. The variance

along the $x$ and $y$ dimension was 100 and the covariance was 0 for all categories.

### Procedure

The procedure was identical to that from Experiment 1.

## Results

### Accuracy analyses

*Pre-change performance* The learning curves for all three conditions across the three pre- and three post-change blocks are presented in Fig. 5a. To verify that there were no differences in pre-change performance, we conducted a three condition × three pre-change block ANOVA. The main effect of block was significant [$F(2, 188) = 77.63$, $p < .001$, $\eta^2 = .452$], whereas the main effect of condition [$F(2, 94) = 1.79$, $p = .17$, $\eta^2 = .037$] and the interaction were non-significant [$F(4, 188) = .52$, ns, $\eta^2 = .011$]. Importantly, no performance differences were observed across conditions in the final pre-change block [$F(2, 94) = .54$, ns, $\eta^2 = .011$]. Thus, pre-change training performance was equated across conditions.

*Cost* The performance cost data are displayed in Fig. 5b. The cost was significantly larger than zero in the Category Label [$t(36) = 8.83$, $p < .001$] and Response Location [$t(40) = 4.37$, $p < .01$] conditions, but did not differ signifi-

cantly from zero in the Control condition [$t(18) = -1.58$, $p = .13$]. There was a main effect of condition on the performance cost [$F(2, 94) = 21.17$, $p < .001$, $\eta^2 = .311$]. In addition, the cost was significantly larger in the Category Label condition than in the Response Location condition [$t(76) = 3.46$, $p < .001$], and than in the Control condition [$t(54) = 6.58$, $p < .001$]. The cost was also significantly larger in the Response Location condition than in the Control condition [$t(58) = 3.72$, $p < .001$].

Examining performance during the first 50-trials yielded the same pattern of results. The cost proportions were $-.026$, .208, and .119 in the Control, Category Label, and Response Location conditions, respectively. The cost was significantly larger than zero in the Category Label [$t(36) = 9.78$, $p < .001$] and Response Location [$t(40) = 5.29$, $p < .001$] conditions, but not in the Control condition [$t(18) = 1.18$, ns]. There was a main effect of condition on the performance cost [$F(2, 94) = 20.39$, $p < .001$, $\eta^2 = .303$]. In addition, the cost was significantly larger in the Category Label than in the Response Location condition [$t(76) = 2.88$, $p < .01$], and than in the Control condition [$t(54) = 6.95$, $p < .001$]. The cost was also significantly larger in the Response Location condition than in the Control condition [$t(58) = 3.99$, $p < .001$].

*Recovery* The recovery data are displayed in Fig. 5c. There was significant recovery both in the Category Label [$t(36) = 8.27$, $p < .001$] and the Response Location [$t(40) = 3.16$, $p < .01$] conditions. In addition, the recovery was significantly larger in the Category Label condition than in the Response Location condition [$t(76) = 2.45$, $p < .05$].

The most important finding is that we observed a larger cost in the Category Label condition than in the Response Location condition, and in this case, the difference was highly significant for both the 50- and 100-trial analyses. This supports a two-association model of information–integration learning. Interestingly, whereas we found no recovery rate difference between the Category Label and Response Location conditions for rule-based categories (Experiment 1), we did find a larger recovery for the Category Label condition with information–integration categories (Experiment 2) suggesting a possible processing dissociation between the rule-based and information–integration classification learning systems. We conduct a more direct comparison of the results from Experiments 1 and 2 in the Discussion section below, but first we briefly discuss the models.

The accuracy-based analyses suggest that the Category Label manipulation led to a larger cost but greater recovery than the Response Location manipulation. However, it remains unclear, whether the manipulations led participants to abandon information–integration decision strategies and

to fall back on rule-based strategies, or whether the manipulations interfered with the implementation of information–integration strategies. To answer these questions, we fit information–integration, rule-based and random responder models to each participant's responses on a block by block basis (Maddox & Ashby, 1993).

The model-based analyses suggested that a large number of participants switched from information–integration strategies during the final pre-change block to non-information–integration strategies during the first 50 trials of the first post-change block in the Category Label and Response Location conditions. However, the use of an information–integration strategy recovered quickly by the second 50 trials of the first post-change block in both conditions. Interestingly, the decrease in the use of information–integration strategies was not associated with a large increase in the proportion of data sets best fit by a rule-based model, but instead was associated with a large increase in the percentage of data sets best fit by the random responder models in both conditions. However, the proportion of data sets best fit by the random responder model remained high in the second 50 trials of the first post-change block for the Category Label condition but not in the Response Location condition. It is important to note that a good fit of the random responder model can be interpreted in one of two ways. One possibility is that participants are truly random in their responding and have completely abandoned the use of a consistent response strategy. A second, and most likely possibility, is that participants are trying multiple strategies in an attempt to find one that "works".

Brief discussion of Experiments 1 and 2

Experiments 1 and 2 used identical Category Label and Response Location manipulations to affect the category label and response location associations of processing differentially. The critical difference between the studies was in the nature of the category learning task. Experiment 1 used a four-category, conjunctive rule-based task, whereas Experiment 2 used a four-category, information–integration task derived by rotating the stimuli from the rule-based categories by 45° in the stimulus space. In this section, we conduct some direct statistical comparisons across the two studies to determine whether the effects of the category label and response location manipulations interact with the nature of the category structure. These findings should be interpreted with some caution because they involve across experiment comparisons. Even so, they provide some evidence to suggest that processing dissociations exist between the rule-based and information–integration systems. As a starting point we compared final block pre-change performance, by conducting a two category structure × three condition ANOVA. The main effects of

category structure [$F(1, 192) = 2.668$, $p = .104$, $\eta^2 = .014$] and condition [$F(2, 192) = .326$, ns, $\eta^2 = .003$] and the interaction [$F(2, 192) = .133$, ns, $\eta^2 = .001$] were all non-significant. Thus, any performance differences that emerge across Experiments 1 and 2 cannot be attributed to differences in pre-change performance.

Next, we compared the performance costs by conducting a two category structure × two condition ANOVA. The main effect of category structure was significant [$F(1, 155) = 21.44$, $p < .001$, $\eta^2 = .121$] and suggested that the performance cost was smaller in the rule-based condition (.040) than in the information–integration condition (.104). The main effect of condition was also significant [$F(1, 155) = 19.403$, $p < .001$, $\eta^2 = .111$] and suggested that the cost associated with the category label manipulation (.102) was larger than that associated with the response location manipulation (.042). The interaction was non-significant [$F(1, 155) = 1.273$, ns, $\eta^2 = .008$]. Finally, we compared the recovery rates by conducting a two category structure × two condition ANOVA. The main effect of category structure was significant [$F(1, 155) = 16.133$, $p < .001$, $\eta^2 = .094$] and suggested that the recovery rate was smaller in the rule-based condition (.040) than in the information–integration condition (.110). The main effect of condition was also significant [$F(1, 155) = 4.954$, $p < .05$, $\eta^2 = .031$] and suggested that the recovery rate associated with the category label manipulation (.094) was larger than that associated with the response location manipulation (.055). Perhaps most interestingly, the interaction was significant [$F(1, 155) = 3.772$, $p = .054$, $\eta^2 = .024$]. As suggested by the results from Experiments 1 and 2, the interaction was characterized by significantly larger recovery in the category label condition than in the response location condition for information–integration categories, but not for rule-based categories.

Taken together, these results suggest that the two associations of rule-based learning identified by Kruschke (1996)—category label and response location—also apply to information–integration learning. In addition, they suggest that the costs associated with the category label and response location manipulations does not interact with the nature of the categorization problem (i.e., rule-based or information–integration). However, there is clear evidence that the recovery rate associated with the category label and response location manipulations *does* interact with the nature of the categorization problem. Specifically, the recovery rate being associated with the category label manipulation is greater than that associated with the response label manipulation for information–integration categories, but not for rule-based categories.

One potential criticism of Experiments 1 and 2 is that it is logically possible that participants in both experimental conditions could learn the transfer categories via a cognitive remapping of the response keys. For example, in all pre-change conditions, the stimuli in Fig. 2 denoted by the open squares are associated with the "Z" key. During post-change training in the Category Label(A) and Response Location(A) conditions, these same stimuli are now associated with the "W" key. It is possible that participants become consciously aware of this change and explicitly remap the open square stimuli with the "W" key. There are at least two arguments against this criticism. First, we observed a different pattern of results for the Category Label and Response Location conditions in both experiments. It is difficult to imagine how a cognitive remapping in both conditions would lead to such different results across the two conditions. Second, there is an overwhelming body of evidence to suggest that information–integration strategies are not available to conscious awareness because they involve the procedural-learning system (Ashby & Casale, 2003; Ashby & Ennis, 2006; Ashby & Maddox, 2005; Maddox & Ashby, 2004). Despite the evidence against a cognitive remapping hypothesis, we decided to run a third experiment that provides a second examination of the effects of these two-category shift manipulations on information–integration classification learning.

## Experiment 3

In Experiment 3, participants learned a two-category information–integration task using stimuli composed of circular sine-wave gratings that varied across trials in bar width and bar orientation (e.g., as in Fig. 1). After pre-change training was complete, the response keys remained the same in the Control condition, or the response keys were switched (i.e., the A key became the B key and vice versa) in the Response Location condition. Note that these procedures mimic those used in Experiments 1 and 2. A scatterplot of the stimuli used during pre-change training, and post-change training in the Control and Response Location conditions are displayed in Fig. 6a. The third condition is referred to as the Rotation condition. After training was complete in the Rotation condition, the category structures were rotated 90° in stimulus space (i.e., in bar width, orientation space). A scatterplot of the stimuli in the Rotation condition are displayed in Fig. 6b. Note that the rotation affects the stimulus-to-category label associations, but in a way that is different from that used in the Category Label conditions of Experiments 1 and 2, and cannot be solved by a simple cognitive remapping. Notice also that this is a variant of the partial shift procedure used by Sanders (1971) and Wills et al. (2006) because some pre-change stimuli remain in the same category, whereas others do not.

Note that in the Response Location condition, every stimulus requires a new response during transfer. In the
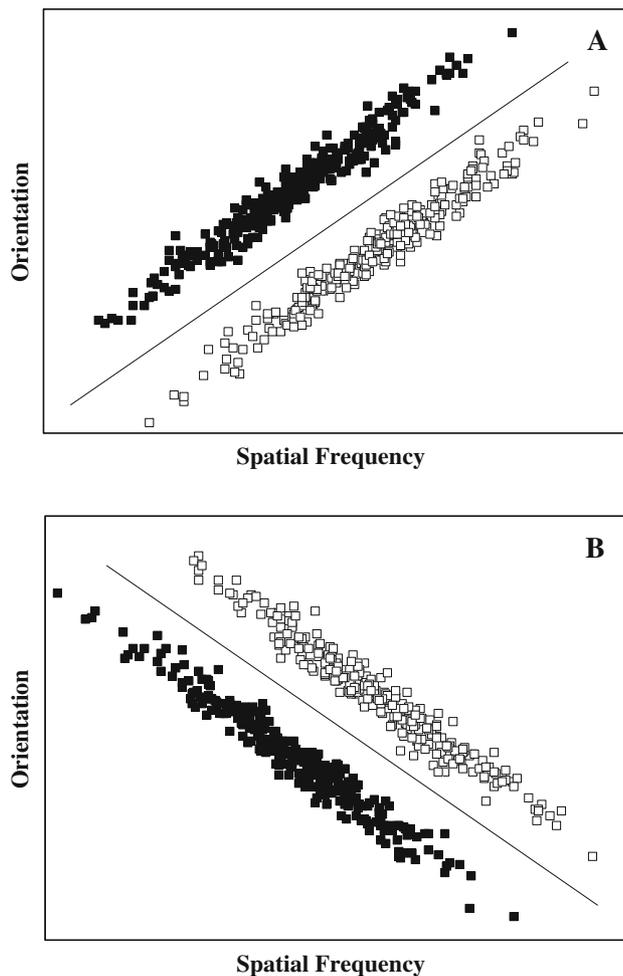
**Fig. 6** Category structures used for Experiment 3. The optimal decision bound is indicated by the *diagonal line*. Category A stimuli are denoted by the open squares and category B stimuli are denoted by the filled squares. **a** Categories used during training in all conditions and during transfer in the Control and Response Location conditions. **b** Categories used during transfer in the Rotation condition

Rotation condition, on the other hand, some stimuli require a new response, some stimuli require the originally trained response, and some stimuli are novel. If a cognitive remapping explains the results from Experiment 2, then it would predict that a smaller performance cost should be observed in the Rotation condition then in the Response Location condition. On the other hand, if manipulations that break the learned association between stimuli and category labels lead to larger costs as observed in Experiment 2, then we would predict that a larger performance cost should be observed in the Rotation condition then in the Response Location condition.

The Rotation and Response Location conditions will not be run with rule-based categories, as these already exist in the literature. For example, Ashby et al. (2003) used a one-dimensional rule-based task and found that a response location switch led to no performance cost, whereas Ell (2003)

found that a 90° rotation led to a large performance cost. A 90° rotation of a one-dimensional rule-based task is equivalent to an extra-dimensional shift (i.e., a previously irrelevant dimension becomes relevant), and many studies have reported that transfer costs are associated with such shifts (Downes et al., 1989; Ell, 2003; Owen, Roberts, Polkey, Sahakian, & Robbins, 1991; Owen et al., 1993).

Method

*Participants*

Seventy-nine participants completed the study and received course credit for their participation. All participants had normal or corrected to normal vision. Each participant served in one condition. To ensure that only participants who learned the initial category structures were included in the transfer performance analyses, a learning criterion of 80% correct in any of the ten pre-change blocks was applied. Of the seventy-nine participants, 53 met the learning criterion (Control $N = 16$; Rotation $N = 19$; Response Location $N = 18$).[1]

*Stimuli and stimulus generation*

Each exemplar in Experiment 1 was a circular sine-wave grating of the type shown in Fig. 1. The stimuli varied across trials on bar width and bar orientation. Category exemplars for the control and the response location conditions were generated by defining bivariate normal distributions along the two stimulus dimensions with mean $x$ values of 300 and 400 and mean $y$ values of 400 and 300 for categories A and B, respectively. The $x$ and $y$ variances were 8,000 with a covariance of 7,800 for both categories. Three hundred random samples were drawn from each distribution (600 total) and were linearly transformed so that the sample mean vector and sample variance–covariance matrix exactly equaled the population mean vector and variance–covariance matrix. Each random sample $(x, y)$ was converted to a stimulus by deriving the orientation (in degrees counterclockwise from horizontal) as $o = x - 30$, and spatial frequency (in cycles per degree) as $f = (y \times .001) + .01$. These scaling factors were chosen to roughly equate the salience of each dimension. The complete set of 600 stimuli is displayed in Fig. 6a. Five hundred of these stimuli were randomly sampled and were used during the training phase of all three conditions. The remaining 100 were used during the transfer phase of the Control and Response Location conditions. The stimuli for the Rotation

---

[1] Several more and less stringent inclusion criteria were examined and the same qualitative pattern in the results emerged. In fact, the same qualitative pattern held when all participants were included.

transfer condition were generated by randomly sampling 100 stimuli from the full complement of 600 and then rotating those by 90° clockwise. The rotated stimuli are shown in Fig. 6b. The optimal decision bound in both panels is denoted by the diagonal line. A participant responding "A" to any exemplar above the optimal categorization bound and "B" to any exemplar below the bound would achieve 100% correct; the best one-dimensional rule (i.e., where the participant based their categorization on a single dimension) would achieve approximately 70% correct.

### Procedure

Each participant was randomly assigned to one of three experimental conditions: Control, Rotation, and Response Location. The experimental session consisted of 10, 50-trial pre-change training blocks, followed by 2, 50-trial post-change transfer blocks, with a participant-controlled rest period after each block. Participants in all three conditions were given response instructions at the beginning of the experiment prior to the first pre-change block, and again immediately prior to the first post-change transfer block. Participants were told that they were to categorize disk patterns on the basis of the orientation and bar width of the stimulus presented on each trial into one of two categories by pressing one button labeled "A" or the other button labeled "B". In all three conditions, participants were given 10, 50-trial blocks of pre-change training with a 5-s response deadline. On each trial, a single stimulus was presented at fixation and the participant was instructed to make a categorization response by pressing one of two response keys (labeled "A" or "B") with their index fingers. The fixation point was presented on a gray background at the center of the screen for 500 ms followed by the stimulus, which was response terminated. Feedback was immediate, and took the form of a 500 Hz tone for 500 ms for correct responses or a 200 Hz tone for 500 ms for incorrect responses. After the feedback interval, a blank screen was presented for 1,000 ms, at which point the next trial began.

Following completion of the 500 pre-change training trials in the Control condition, participants were given 2, 50-trial blocks of post-change transfer with a 1.5-s response deadline. The response deadline was included to ensure that the participants could not overcome any interference by simply inhibiting their initial response. If a response was not made before the response deadline, the participant received a message to speed up his or her response, and the trial was discarded from subsequent analyses. On average, this occurred on 3 of the 100 transfer trials. Post-change transfer in the Response Location condition was identical to that in the Control condition with the exception that the buttons associated with categories A and B were reversed so that the button labeled "A" was now associated with

response "B" and vice versa, thereby switching the response locations. Post-change transfer in the Rotation condition was identical to that in the Control condition with the exception that 100 stimuli from the rotated category structure were utilized (see Fig. 6b). Participants in both the Response Location and Rotation conditions were given instructions that during the post-change transfer phase the stimuli or the response mappings would change, but were not told which would change.
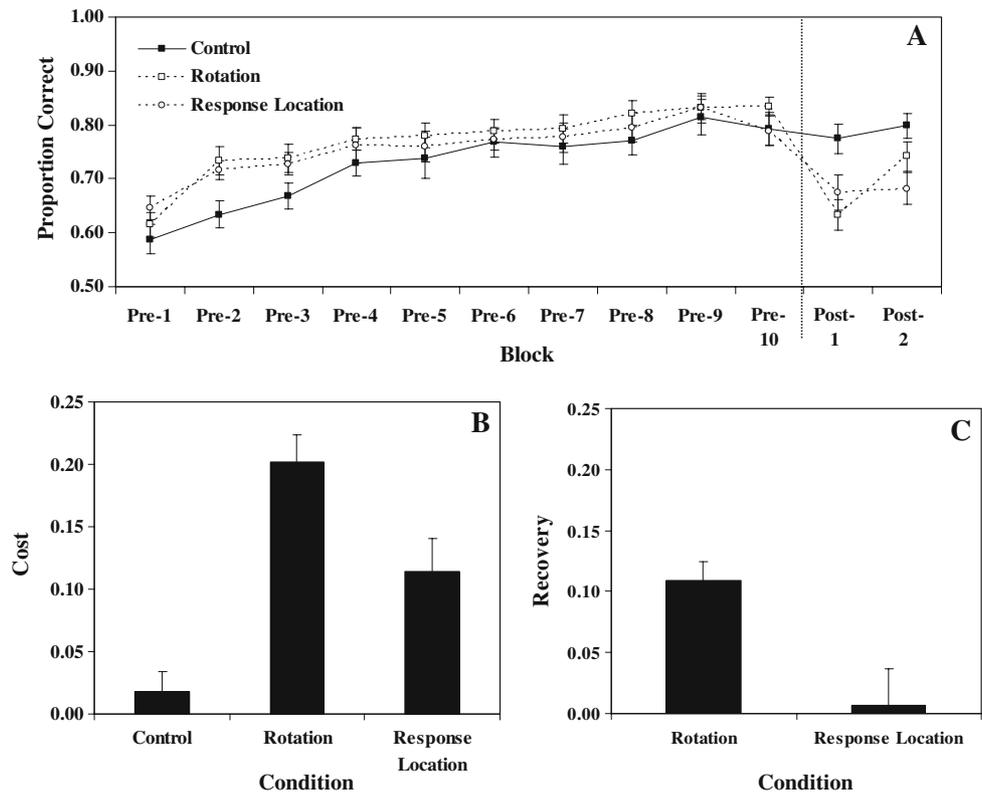
### Results

#### Accuracy analyses

*Pre-change performance*   The learning curves for all three conditions across the ten pre- and two post-change blocks are presented in Fig. 7a. To verify that there were no differences in pre-change performance, we conducted a three condition × ten pre-change block ANOVA. The main effect of block was significant [$F(9, 450) = 31.64$, $p < .001$, $\eta^2 = .388$], whereas the main effect of condition [$F(2, 50) = 1.52$, $p = .23$, $\eta^2 = .057$] and the interaction were non-significant [$F(18, 450) = .768$, ns, $\eta^2 = .030$]. In addition, there was no effect of condition in the tenth pre-change block [$F(2, 50) = 1.12$, ns, $\eta^2 = .043$]. Thus, pre-change training performance was equal across conditions.

*Cost*   The performance cost data are displayed in Fig. 7b. As expected, there was no cost in the Control condition [$t(15) = 1.09$, $p > .05$], but the cost was significantly larger than zero in the Rotation [$t(18) = 9.312$, $p < .001$] and Response Location conditions [$t(17) = 4.45$, $p < .001$]. The main effect of condition on the performance cost [$F(2, 50) = 17.21$, $p < .001$, $\eta^2 = .408$]. In addition, the cost was significantly larger in the Rotation condition than in the Response Location condition [$t(35) = 2.61$, $p = .013$], and than in the Control condition [$t(33) = 6.57$, $p < .001$]. The cost was also significantly larger in the Response Location condition than in the Control condition [$t(32) = 3.08$, $p < .01$].

*Recovery*   The recovery data are displayed in Fig. 7c. There was no significant recovery in the Response Location condition [$t(17) < 1.0$], but recovery was significant in the Rotation condition [$t(18) = 3.66$, $p < .01$]. In addition, recovery was significantly larger in the Rotation condition than in the Response Location condition [$t(35) = 2.67$, $p = .011$]. Importantly, the lack of recovery in the Response Location condition cannot be interpreted as a ceiling effect since there was considerable room for a performance improvement.

The most important finding is that the cost and recovery rates were larger in the Rotation condition than in the

**Fig. 7** **a** Proportion correct (averaged across participants) from Experiment 3. **b** Cost determined by subtracting post-change block 1 performance from pre-change block 10 performance. **c** Recovery determined by subtracting post-change block 1 performance from post-change block 2 performance. Standard *error bars* included



Response Location condition. This replicates the pattern observed in the Category Label and Response Location conditions of Experiment 2, and suggests that a cognitive remapping explanation for the Experiment 2 results is unlikely. Thus, across two different experimental manipulations of the category label association relative to the response location association, and with two different stimulus sets, we find the same pattern, with larger costs for manipulations that disrupt the category label than for manipulations that disrupt the category-response mapping. This provides further evidence that these two learned associations, originally proposed with respect to rule-based classification, apply to information–integration classification, and suggests that the adverse effects are larger, but the recovery faster when the manipulation affects the category label association of information–integration classification learning.

The accuracy-based analyses suggest that the Rotation condition led to a larger cost but greater recovery than the Response Location condition. As with Experiment 2, it is important to determine whether the manipulations led participants to abandon information–integration decision strategies, and to fall back on rule-based strategies, or whether the manipulations interfered with the implementation of information–integration strategies. To address this issue we fit models to the data from the final pre-change block and both post-change blocks.

The model-based analyses indicate that a large number of participants switched strategy types when instructed that the categories were going to change. However, it appears that switching back to a response strategy of the same type as the optimal classifier (i.e., information–integration) may have been more difficult when the category structures were rotated than when the response locations were switched. In other words, more participants in the Response Location condition were able to relearn an information–integration response strategy than participants in the Rotation condition. It is also worth mentioning that in all but one case for which a rule-based strategy fit best, the best-fitting model assumed a simple one-dimensional rule.

The accuracy and modeling data support the hypothesis that rotating the categories interfered enough with information–integration categorization that participants abandoned information–integration strategies in the Rotation condition in favor of rule-based strategies and that they persisted with this rule use to a greater extent than did participants in the Response Location condition. Note that this result is also consistent with the greater recovery rate seen in the Rotation condition. This follows because with two categories, rule-based learning is faster than information–integration learning, and in the present experiment a one-dimensional rule can perform reasonably well (i.e., approximately 70% correct). As a result, participants who used explicit rules during transfer could quickly increase their accuracy. In

fact, an accuracy analysis shows that participants in the Rotation condition who used rule-based strategies increased their accuracy from 63% in the first transfer block to 71% correct in the final transfer block.

Discussion

Rotating the stimuli and switching the response keys each impaired performance, but the cost and recovery were larger in the Rotation condition than the Response Location condition. These results mirror those from Experiment 2, and suggest that there are two associations involved in information–integration learning, and that manipulations of the category label association lead to larger costs but faster recovery then manipulations of the label-to-response association. The Experiment 3 results also provide evidence against a cognitive remapping explanation of the Experiment 1 and 2 results. In the Response Location condition the correct response changed for every region of stimulus space, whereas in the Rotation condition this was true for only part of the stimulus space. Thus, a model that assumes a cognitive remapping would predict that all responses must be relearned in the Response Location condition, but only half the responses need relearning in the Rotation condition. For this reason, a cognitive remapping model could account for *smaller* costs in the Rotation condition, but would be incompatible with the present results in which the cost was larger in the Rotation condition.

In both conditions, the experimental manipulation caused a decrease in the use of information–integration strategies and an increase in the use of rule-based strategies. The larger recovery in the Rotation condition was not associated with an increase in the use of information–integration strategies as one might expect. Instead, rule-based strategies continued to dominate in the second transfer block of the Rotation condition, whereas information–integration strategy use returned in the Response Location condition, although in both cases at least 50% of the participants used a rule-based strategy in the second transfer block.

**General discussion**

Understanding how people adapt to novel and changing environments is an important focus of psychological research. One of the most popular paradigms for investigating these processes is to train people on rule-based classification tasks that can be learned via explicit reasoning, and then to examine the performance costs and recovery rates associated with various shifts in the nature of the problem. Research examining the effects of category shifts suggests that there are at least two associations

involved in rule-based classification learning: a stimulus-to-label association that learns to map groups of stimuli with a category label, and a label-to-response association that learns to map the category labels with responses (Goldstone & Steyvers, 2001; Kendler & Kendler, 1962; Kruschke, 1996).

The existence of these separate associations was confirmed in a conjunctive, rule-based learning task and was extended to two different information–integration category learning problems. Experimental conditions that either disrupted the stimulus-to-category label or the category label-to-response location mapping were examined. Disruptions of the stimulus-to-category label association led to a larger cost and greater recovery than disruptions of the category label-to-response location association for information–integration categories. Disruptions of the stimulus-to-category label association led to a larger cost but equivalent recovery than disruptions of the category label-to-response location association for rule-based categories. This suggests some similarities across rule-based and information–integration categories (e.g., in the effects on costs), but also some important differences (e.g., in the effects on recovery rates).

This study advances the field in a number of directions. First, stimulus-to-category label manipulations have not been studied extensively in information–integration category learning. These data suggest that manipulations of this sort lead to robust initial performance costs, but large rates of recovery. Second, these data build upon initial work by Wills et al. (2006) who compared full and partial reversal conditions in family resemblance categories. They found evidence to suggest that the stimulus-to-category label association might actually be decomposable into two sub-associations. This possibility is discussed below. Finally, this work extends that conducted by Ashby et al. (2003; see also Maddox et al., 2007) to a complex four-category information–integration problem that includes sufficient post-change training to examine recovery rates. We turn now to a brief discussion of a number of important topics.

Two learned associations in information–integration learning: theoretical and neurobiological implications

Of the many theories of category learning, only one posits a neurobiological locus of information–integration category learning. The COmpetition between Verbal and Implicit Systems (COVIS; Ashby & Waldron, 1999; Ashby et al., 1998) model assumes that only a single association is involved in information–integration category learning that is essentially equivalent to the stimulus-to-label association proposed here. In COVIS, associations between stimuli and category labels are learned via changes in synaptic strength between pyramidal cells in visual association areas and

medium spiny cells in the striatum. A large set of diverse results support this general model (for a review see, e.g., Ashby & Ennis, 2006), including single-cell recording studies in monkeys showing that striatal medium spiny cells develop category-specific responses after extensive categorization training (Merchant, Zainos, Hernandez, Salinas, & Romo, 1997; Romo, Merchant, Ruiz, Crespo, & Zainos, 1995; Romo, Merchant, Zainos, & Hernandez, 1997). Some of the characteristics associated with the neural architecture of the COVIS stimulus-to-label association are roughly consistent with the empirical findings reported here. For example, breaking the associations between the stimuli and category labels should effectively return the synaptic strengths to baseline levels leading the procedural system to relearn from scratch. This would lead to a large performance cost, and a rate of recovery similar to that observed in initial learning as if a new classification task was being performed.

Because COVIS postulates no learning after the cortical-striatal synapses, it would predict no qualitative difference between the rotation/category label and response location manipulations used here. Both manipulations would require cortical-striatal relearning. For this reason, COVIS, in its current form, is not consistent with the present results. COVIS might be extended, however, to include a second learned association that consists of associating a category label with a specific response location. Logically, such learning must be downstream from the site of category label learning, which suggests that plausible sites of response learning could be at synapses in the internal segment of the globus pallidus, the ventral anterior or ventral lateral nuclei of the thalamus, or perhaps within premotor cortex. Each of these brain regions has been implicated in procedural learning and, thus represent plausible loci of such learning (Poldrack et al., 2005). However, at present, we know of no behavioral or neuroscience data that could be used to narrow this search. As such, proposing a neurobiological basis for label-to-response learning must remain a goal of future research. Even so, the current results do offer some insights into the behavioral properties of the label-to-response associations (smaller cost, and weaker recover), and thus should help narrow the field of possible neural loci.

### An alternative two-association model

We interpret the current data as supporting a two-association model that assumes that one association maps stimuli to labels and a second maps labels-to-responses. A reasonable alternative, however, might be a two-association model that assumes a stimulus-to-label association (as we do) but also a stimulus-response (instead of a category label-to-response) association that learns the direct

mapping between the stimulus and the response.[2] One advantage of this model is that it provides a straightforward explanation for why the category label manipulation led to a larger performance cost than the response location manipulation. In short, whereas both the category label and response location manipulations broke the learned stimulus-response association, only the category label manipulation also broke the learned stimulus-to-category label association. Thus, two associations were broken by the category label manipulation, whereas only one was broken by the response location manipulation. Note that this is *not* a viable two-association model of rule-based classification, because it predicts a performance cost anytime that the stimulus-response association is broken. As outlined earlier, Ashby et al. (2003) reported that a button switch did not adversely affect two-category rule-based performance, whereas it did affect two-category information–integration performance. Even so, it does provide a viable model of information–integration classification. Thus, although these data clearly support a two-association model of information–integration classification, a determination of the exact nature of the associations awaits future research.

### Future directions

There are a number of exciting directions to take this work. One that is particularly promising is to explore in greater detail the possibility that rule-based and information–integration classification are characterized by three, as opposed to two, learned associations. Whereas the current research supports the existence of a stimulus-to-label and a label-to-response (or perhaps stimulus-response) association, the work of Sanders (1971) and Wills et al. (2006), using an optional shift paradigm suggests that the stimulus-to-label association might actually be composed of two separate associations: a stimulus-to-category representation association and a category-representation-to-category label association. Unfortunately, the many differences in procedures and stimuli make a definitive conclusion premature. Future research should attempt to provide evidence for all three associations *within* a single experiment.

A second line of work should focus on the processing characteristics associated with each association. Here an application of a processing dissociation approach would be highly useful. Over the past decade or so, Ashby, Maddox, and their colleagues used a process dissociation approach to provide convincing evidence that rule-based category learning is mediated by an explicit hypothesis-testing system, whereas information–integration category learning is mediated by an implicit procedural-learning-based system.

---

[2] We thank an anonymous reviewer for suggesting this alternative.

To achieve this goal, experimental manipulations were introduced that affected the explicit system but not the implicit system (or vice versa) and the effects of these manipulations on rule-based and information–integration category learning were examined. For example, changes to the nature and timing of the feedback disrupted information–integration, but not rule-based category learning (Ashby, Maddox, & Bohil, 2002; Maddox & Ing, 2005; Maddox, Ashby, & Bohil 2003; Maddox, Love, Glass, & Filoteo, 2008), whereas increasing the working memory load disrupted rule-based, but not information–integration category learning (Waldron & Ashby, 2001; Zeithamova and Maddox 2006, 2007). The coherence of the categories has also been shown to affect information–integration but not rule-based category learning (Maddox, Filoteo, & Lauritzen, 2007; Maddox, Filoteo, Lauritzen, Connally, & Hejl, 2005). Applying this same process dissociation approach to the various associations involved in rule-based and information–integration classification learning will provide many important insights into the nature of processing in each association.

A third line of work should focus on the pre- to post-change transition instructions. In all experiments participants were informed that "some" aspect of the task changed (e.g., the assignment of categories to buttons had changed). This is in contrast to many shift studies that provide the participants with no information. One prediction is that the removal of the transition instructions will have no effect on the magnitude of the effect in the implicit, information–integration task, but will have an effect in the explicit, rule-based task.

Another interesting approach would be to look at category label and response location conditions that introduced new category labels and new response locations. In our studies the same category labels and response locations were used pre- and post-change, but the stimulus-to-category label and category label-to-response location assignments were changed. We took this approach because it is the most common in the extant literature. Even so, future work should examine cases in which the category labels and response locations during the post-change phase are novel.

## Conclusions

The existence of separate stimulus-to-label and label-to-response associations, hypothesized in previous work (Goldstone & Steyvers, 2001; Kendler & Kendler, 1962; Kruschke, 1996), was confirmed in conjunctive, rule-based learning and was extended to information–integration category learning. Separate conditions that either disrupted the mapping from stimulus-to-category label or disrupted the

mapping from category label-to-response location were examined. Both manipulations led to significant performance costs in information–integration learning, but disrupting the stimulus-to-category label mapping led to a significantly larger cost than disrupting the category label-to-response mapping. Significantly larger costs associated with disrupting the stimulus-to-label mapping relative to disrupting the label-to-response mapping were also observed in rule-based learning. In addition, recovery was greater when the stimulus-to-label mapping was broken in information–integration learning. With rule-based categories, a qualitatively different performance pattern emerged in which no difference in the magnitude of the recoveries was observed. These results provide strong behavioral evidence that information–integration category learning is mediated by separate stimulus-to-label and label-to-response learning associations.

## References

Ashby, F. G., & Casale, M. (2003). The cognitive neuroscience of implicit category learning. In L. Jimenez (Ed.), *Attention and implicit learning* (pp. 108–141). Amsterdam: John Benjamins Publishing Company.

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *The Psychology of Learning and Motivation, 47*, 1–36.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

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

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., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory and Cognition, 30*(5), 666–677.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory and Cognition, 31*(7), 1114–1125.

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

Buss, A. H. (1953). Rigidity as a function of absolute and relational shifts in the learning of successive discriminations. *Journal of Experimental Psychology, 45*(3), 153–156.

Buss, A. H., & Buss, E. H. (1956). The effect of verbal reinforcement combinations on conceptual learning. *Journal of Experimental Psychology, 52*(5), 283–287.

Downes, J. J., Roberts, A. C., Sahakian, B. J., Evenden, J. L., Morris, R. G., & Robbins, T. W. (1989). Impaired extra-dimensional shift

performance in medicated and unmedicated Parkinson's disease: Evidence for a specific attentional dysfunction. *Neuropsychologia, 27*, 1239–1243.

Ell, S. W. (2003). Selection and switching in rule-based category learning. Unpublished Dissertation, University of California, Santa Barbara.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Filoteo, J. V., Maddox, W. T., Simmons, A. N., Ing, A. D., Cagigas, X. E., Matthews, S., et al. (2005). Cortical and subcortical brain regions involved in rule-based category learning. *Neuroreport, 16*(2), 111–115.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*(1), 116–139.

Kendler, H. H., & Kendler, T. S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review, 69*, 1–16.

Kendler, H. H., & Kendler, T. S. (1968). Mediation and conceptual behavior. In K. W. S. J. T. Spence (Ed.), The psychology of learning and motivation (Vol. 2, pp. 197–244). New York: Academic Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science, 8*(2), 225–247.

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

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes, 66*(3), 309–332.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(1), 100–107.

Maddox, W. T., 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*(4), 650–662.

Maddox, W. T., Filoteo, J. V., Lauritzen, J. S., Connally, E., & Hejl, K. D. (2005). Discontinuous categories affect information-integration but not rule-based category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(4), 654–669.

Maddox, W. T., Filoteo, J. V., & Lauritzen, J. S. (2007a). Within-category discontinuity interacts with verbal rule complexity in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 33*(1), 197–218.

Maddox, W. T., Lauritzen, J. S., & Ing, A. D. (2007b). Cognitive complexity effects in perceptual classification are dissociable. *Memory and Cognition, 35*(5), 885–894.

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*(2), 578–589.

Merchant, H., Zainos, A., Hernandez, A., Salinas, E., & Romo, R. (1997). Functional properties of primate putamen neurons during the categorization of tactile stimuli. *Journal of Neurophysiology, 77*(3), 1132–1154.

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*(3), 676–692.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

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*(2), 279–291.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., et al. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17*(1), 37–43.

Owen, A. M., Roberts, A. C., Polkey, C. E., Sahakian, B. J., & Robbins, T. W. (1991). Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia, 29*(10), 993–1006.

Owen, A. M., Roberts, A. C., Hodges, J. R., Summers, B. A., Polkey, C. E., & Robbins, T. W. (1993). Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain, 116*, 1159–1175.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews, 32*(2), 197–205.

Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., et al. (2001). Interactive memory systems in the human brain. *Nature, 414*(6863), 546–550.

Poldrack, R. A., Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y., et al. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience, 25*(22), 5356–5364.

Robbins, T. W. (2007). Shifting and stopping: fronto-striatal substrates, neurochemical modulation and clinical implications. *Philosophical Transactions of the Royal Society of London: B Biological Sciences, 362*(1481), 917–932.

Romo, R., Merchant, H., Ruiz, S., Crespo, P., & Zainos, A. (1995). Neuronal activity of primate putamen during categorical perception of somaesthetic stimuli. *Neuroreport, 6*(7), 1013–1017.

Romo, R., Merchant, H., Zainos, A., & Hernandez, A. (1997). Categorical perception of somesthetic stimuli: psychophysical measurements correlated with neuronal events in primate medial premotor cortex. *Cerebral Cortex, 7*(4), 317–326.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative and Physiological Psychology, 74*, 192–202.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews, 32*(2), 265–278.

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience, 25*(11), 2941–2951.

Seger, C. A., & Cincotta, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex, 16*(11), 1546–1555.

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

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

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin and Review, 8*(1), 168–176.

Wills, A. J., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory and Cognition, 34*(1), 17–27.

Wolff, L. L. (1967). Concept-shift and discrimination-reversal learning in humans. *Psychological Bulletin, 68*, 369–408.

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

Zeithamova, D., & Maddox, W. T. (2007). The role of visuo-spatial and verbal working memory in perceptual category learning. *Memory and Cognition, 35*(6), 1380–1398.