



Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty

F. Gregory Ashby¹ · J. David Smith² · Luke A. Rosedahl³

Published online: 16 December 2019
© The Psychonomic Society, Inc. 2019

Abstract

In rule-based (RB) category-learning tasks, the optimal strategy is a simple explicit rule, whereas in information-integration (II) tasks, the optimal strategy is impossible to describe verbally. Many studies have reported qualitative dissociations between training and performance in RB and II tasks. Virtually all of these studies were testing predictions of the dual-systems model of category learning called COVIS. The most prominent alternative account to COVIS is that humans have one learning system that is used in all tasks, and that the observed dissociations occur because the II task is more difficult than the RB task. This article describes the first attempt to test this difficulty hypothesis against anything more than a single set of data. First, two novel predictions are derived that discriminate between the difficulty and multiple-systems hypotheses. Next, these predictions are tested against a wide variety of published categorization data. Overall, the results overwhelmingly reject the difficulty hypothesis and instead strongly favor the multiple-systems account of the many RB versus II dissociations.

Keywords Categorization · Task difficulty · COVIS · AlexNet

Introduction

A large literature compares performance in rule-based (RB) and information-integration (II) categorization tasks. Typical examples of such tasks are shown in Fig. 1. Each task includes two categories of stimuli that vary across trials on two stimulus dimensions. In the case of Fig. 1, these stimuli are circular sine-wave gratings (e.g., Gabor patches) that vary across trials in bar width and bar orientation. In standard applications, stimuli are presented one at a time, participants assign each stimulus to a category by pressing a response key, and feedback is given after each response (i.e., correct vs. incorrect). Participants are told that there are two categories of stimuli and that their task is to use the feedback to learn how to assign

each stimulus to its correct category. Critically, they are given no prior information about the structure of the categories.

In RB tasks, the optimal strategy is a relatively simple explicit rule that can be described using Boolean algebra (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). In the simplest variant, illustrated in Fig. 1, only one dimension is relevant (bar width), and the task is to discover this dimension and then map the different dimensional values to the relevant categories. In II tasks, accuracy is maximized only if information from two or more incommensurable stimulus dimensions is integrated perceptually at a pre-decisional stage (Ashby & Gott, 1988). In most cases, the optimal strategy in II tasks is difficult or impossible to describe verbally (Ashby et al., 1998). Verbal rules may be (and sometimes are) applied but they lead to poor performance. The categories shown in the two Fig. 1 tasks are simple rotations of each other in stimulus space (i.e., by 45°), and so are exactly equated on many category-separation statistics. They differ only in whether the optimal decision bound is horizontal (RB condition) or diagonal (II condition). Despite their similarity, many studies have shown that during early learning, human accuracy is much higher in the RB task than the II task (i.e., given the same amount of training). The RB advantage lasts for at least a few thousand trials, but after that, performance asymptotes at the same high levels of accuracy in both tasks (e.g., Hélié, Waldschmidt, & Ashby, 2010).

✉ F. Gregory Ashby
fgashby@ucsb.edu

¹ Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA 93106, USA

² Department of Psychology, Georgia State University, Atlanta, GA, USA

³ Interdepartmental Graduate Program in Dynamical Neuroscience, University of California, Santa Barbara, CA, USA

Many studies have compared performance in RB and II tasks, like those shown in Fig. 1, across a variety of different training and testing conditions. These studies have varied the methods and order of stimulus presentation, the nature and timing of feedback, the quality of the training conditions, and the generalizability of the knowledge acquired during training. This line of research has produced scores of articles that have reported somewhere between 25 and 30 qualitative dissociations between training and performance in RB and II tasks¹ (for a review of many of these, see Ashby & Valentin, 2017).

Virtually all of these studies were testing predictions of the dual-systems model of category learning called COVIS (Ashby et al., 1998; Ashby & Valentin, 2017; Ashby & Waldron, 1999). Briefly, COVIS assumes that humans learn categories in at least two qualitatively different ways. An executive attentional system uses working memory to learn explicit rules, whereas a procedural system uses dopamine-mediated reinforcement learning when perceptual similarity determines category membership and the optimal strategy is difficult or impossible to describe verbally. The explicit, rule-learning system is assumed to dominate in RB tasks, whereas the procedural learning system dominates in II tasks. COVIS accounts for the many RB versus II dissociations by assuming that the training and testing conditions that were manipulated affected the two systems differently. In fact, this approach allows COVIS to simultaneously account for all of the reported dissociations *a priori* – that is, COVIS predicted that a dissociation in the opposite direction was impossible. For example, COVIS predicts that a simultaneous dual task that requires working memory and executive attention must interfere with RB tasks at least as much as II tasks, because the explicit rule-learning system presumed to dominate in RB tasks uses executive attention and working memory, whereas the procedural-learning system does not. This prediction has been supported in several studies (Crossley, Paul, Roeder, & Ashby, 2016; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006).

No competing or alternative account of all these dissociations has been proposed. However, a number of studies have hypothesized that the results of some individual dissociation could be due to a difficulty or complexity difference between the RB and II tasks (Edmunds, Milton, & Wills, 2015; Le Pelley, Newell, & Nosofsky, 2019; Nosofsky, Stanton, & Zaki, 2005; Zaki & Kleinschmidt, 2014). In particular, each of these studies postulated that there is a single learning system used in all tasks, but that the II task shown in Fig. 1 is more difficult than the RB task, and that this difficulty difference was the cause of the one dissociation that the study examined.

¹ They are qualitative in the sense that the dissociations are not the result of the specific parametric choices of any independent variables. For example, feedback delays impair II learning but have no effect on RB learning whether the delay is 2 s, 5 s, or 10 s (Crossley & Ashby, 2015; Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005).

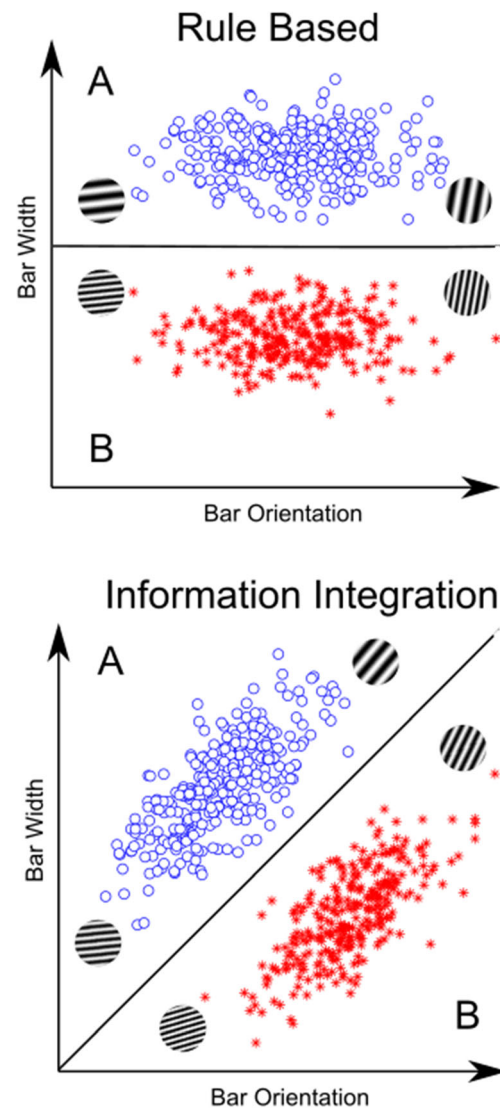


Fig 1 Examples of rule-based (RB) and information-integration (II) category structures. Each stimulus is a sine-wave disc that varies across trials in bar width and bar orientation. For each task, two illustrative Category A and B stimuli are shown. The circles and stars denote the specific values of all stimuli used in each task. In the RB task, only bar width carries diagnostic category information, so the optimal strategy is to respond with a one-dimensional bar-width rule (thick versus thin), while ignoring the orientation dimension. In the II task, both bar width and orientation carry useful but insufficient category information. The optimal strategy requires integrating information from both dimensions in a way that is impossible to describe verbally

This difficulty hypothesis is the most prominent alternative account to COVIS for the many RB versus II dissociations that have been reported. Given this prominence, it is surprising that no attempt has been made to test the difficulty hypothesis against anything more than a single isolated data set. For example, Le Pelley et al. (2019) hypothesized that deferred feedback impairs II learning more than RB learning – as reported by Smith et al. (2014) – because the II task is more difficult or complex than the RB task, but Le Pelley et al. (2019) made no attempt to determine whether this difficulty

account is consistent with the many other empirical RB versus II dissociations that have been reported. To correct this shortcoming in the literature, this article reports the results of the first attempt to test the difficulty hypothesis against a broader range of data. One challenge to testing this hypothesis is that none of the studies arguing that a particular RB-II dissociation was caused by a difficulty difference defined what they mean by difficulty. Even without a rigorous definition, however, it turns out that strong tests are possible.

In the next section, we derive two novel predictions that discriminate between the difficulty and multiple-systems hypotheses. First, we show that the difficulty hypothesis predicts that there must exist some single measure of task difficulty that simultaneously accounts for performance differences in all RB and II tasks. In contrast, the multiple-systems hypothesis predicts that no such measure can exist because comparing difficulty in RB and II tasks is like comparing apples and oranges. Second, we show that the difficulty hypothesis predicts that the ordering of tasks by performance (e.g., final-block accuracy) cannot change with the state of the learner, whereas the multiple-systems hypothesis predicts that such reversals are possible.

In the third section we examine the first test. Specifically, we examine 13 different measures of categorization task difficulty that come from the cognitive-science and machine-learning literatures, and we show that none of these are consistent with performance differences in RB and II tasks. We also show that one of the most powerful and popular deep convolutional neural networks performs identically on the two tasks. Therefore, none of these 13 measures are candidates for the unknown difficulty measure that is required by the difficulty hypothesis, nor is the convolutional neural network that we tested. The fourth section describes empirical evidence that strongly favors the multiple-systems hypothesis on the second of these tests.

In the fifth section we show that even if one leaves difficulty undefined, and just accepts that the II task is more difficult, there are still many empirical phenomena that falsify the hypothesis that the dissociations are due to a difficulty difference. Overall, the evidence reported in this article overwhelmingly rejects the difficulty hypothesis and instead strongly favors the multiple-systems account of the many RB versus II dissociations.

Two novel predictions of the difficulty hypothesis

In this section, we derive two novel and contrasting predictions that differentiate the difficulty and multiple systems hypotheses. The first is that the difficulty hypothesis predicts that there must exist a single difficulty measure that can account for performance differences in all categorization tasks,

whereas the multiple systems hypothesis predicts that such a measure cannot exist. The second differential prediction is that the difficulty hypothesis predicts that the rank order of two tasks by difficulty cannot depend on the state of the learner, whereas the multiple systems hypothesis predicts that at least some difficulty reversals of this type are possible.

Is there a single measure that predicts difficulty in all categorization tasks?

All learning systems struggle under certain experimental conditions and flourish under others. If there is only one learning system, then the same conditions will cause learning to struggle in all tasks. A difficulty measure that is sensitive to these conditions will therefore accurately predict difficulty in all categorization tasks. In contrast, if one system dominates in RB tasks and a different system dominates in II tasks, then conditions that cause learning to struggle will be different in RB and II tasks (else the systems would not be different). As a result, instead of one difficulty measure, two would be required – one that is sensitive to the conditions that make RB learning difficult, and one that is sensitive to the conditions that make II learning difficult. Therefore, one difference between the difficulty and multiple-systems hypotheses is that the difficulty hypothesis requires that a single measure of task difficulty exists that simultaneously accounts for performance differences in all RB and II tasks, whereas the multiple-systems hypothesis requires that constructing such a measure is impossible. The single measure assumed by the difficulty hypothesis might be highly complex in the sense that it could depend on a variety of different category statistics. The only requirement of the difficulty hypothesis is that the same measure must apply to all tasks.

Unfortunately, this prediction of the difficulty hypothesis is almost impossible to refute because doing so would require testing and rejecting an infinite number of different possible measures. Conversely, even if the difficulty hypothesis is true, and there is a single measure of categorization difficulty, it might be difficult to find this measure among the infinite number of alternatives. For this reason, our goals must be somewhat limited. One way to begin, however, is to note that if there is a single valid difficulty measure, then it must accurately predict performance differences among all RB tasks, and also among all II tasks. So one approach is to begin with measures that have successfully predicted difficulty differences among RB tasks and ask whether they also succeed on II tasks, and to examine measures that have been successful with II tasks and ask whether they succeed with RB tasks.

In fact, measures that successfully predict difficulty differences among RB tasks exist, and so do measures that successfully predict difficulty differences among II tasks. In the case of RB categories, the best measure of task difficulty was proposed by Feldman (2000), who hypothesized that difficulty is

determined by the Boolean complexity of the optimal classification rule. He showed that Boolean complexity gave a good account of difficulty differences across 41 different category structures that all had optimal rules that could be described verbally. Rosedahl and Ashby (2019) derived a difficulty measure for II tasks from the procedural-learning model of COVIS, which they called the striatal difficulty measure (SDM). They showed that the SDM accounted for 87% of the variance in final-block accuracy across a wide range of mostly II category-learning data sets, and that this measure provided consistently better predictions than 12 alternative measures. The data sets came from four previously published studies that each included multiple conditions that varied in difficulty. The studies were highly diverse and included experiments with both continuous- and binary-valued stimulus dimensions, a variety of stimulus types, and both linearly and nonlinearly separable categories.

The current question is whether either of these measures, or any others, can simultaneously predict learning difficulty in RB and II tasks. Unfortunately, Boolean complexity of the optimal classification strategy can be ruled out immediately because there is no Boolean algebraic analogue of the diagonal bound in the Fig. 1 II task.² Even so, the SDM can be applied to RB tasks, and many other difficulty measures have been proposed that can be applied to both RB and II tasks. Below we consider 13 measures that are popular in the cognitive-science and machine-learning literatures, and ask whether any of these are at least roughly consistent with performance differences in RB and II tasks. In addition to these measures, we also tested a popular deep convolutional neural network – named AlexNet – on the same RB and II categories (Krizhevsky, Sutskever, & Hinton, 2012). AlexNet famously won the ImageNet Large Scale Visual Recognition Challenge in 2012 by a large margin. This competition required each algorithm to categorize 150,000 photographs of natural objects into one of 1,000 object categories. Although AlexNet may use a different classification strategy than humans, in terms of classification accuracy, it is among the most human-like algorithms available.

Does categorization difficulty depend on the state of the learner?

A fundamental prediction of the multiple-systems hypothesis is that accurate prediction of task difficulty must depend on statistical properties of the categories and on some property of the learner. If there are multiple learning systems and each system struggles under different conditions, then information about which system the learner is using is required to

² To write the Boolean analogue of a categorization strategy, the corresponding decision bound must be a piece-wise linear function that includes only vertical and horizontal line segments.

determine which statistical properties of the categories determine difficulty. In contrast, the difficulty hypothesis predicts that the learner uses the same learning strategy with all categories, and therefore information about the current state of the learner is not necessary to order tasks by difficulty. Knowledge of the learner's state might help predict whether performance will be good or bad, but if the learner struggles more with task A than task B, then this ordering should hold regardless of the current working memory capacity, attentional resources, or motivation of the learner. Therefore, only the categories need to be studied – to extract the statistical properties that determine difficulty. One can see that the difficulty hypothesis is quickly undermined without this constraint. If different task conditions determine difficulty when the learner changes state, then the learner must be using different strategies or processes in different states, in which case the difficulty hypothesis explodes into a perfect version of the multiple-systems hypothesis.

One empirical test between these alternatives would be to ask whether it is possible that performance in some task A is better than performance in task B when the learner is in one state, but that this ordering reverses when the learner is in a different state. The difficulty hypothesis predicts that this scenario is impossible because any state that affects learning should have similar effects on all tasks. For example, reduced motivation should impair performance in all tasks, so the ordering of tasks by difficulty should be preserved across different levels of motivation. However, if performance in task A is mediated by one system and performance in task B is mediated by a different system, then any state of the learner that affects one system more than the other could cause the ordering of tasks by performance to reverse. Thus, a second novel test between the difficulty and multiple-systems hypotheses is to ask whether the ordering of tasks by performance can ever reverse with changes in the state of the learner.

Is there a single measure that predicts difficulty in all categorization tasks?

This section examines the first novel prediction of the difficulty hypothesis – namely, that there must exist a single measure of difficulty based only on statistical properties of the categories that simultaneously accounts for performance differences in all categorization tasks. In contrast, the multiple-systems hypothesis predicts that the best one can do is develop separate quantitative measures of RB and II task difficulty. As mentioned previously, this section examines whether any of 13 different difficulty measures that are popular in the cognitive science and machine learning literatures, or the deep convolutional neural network AlexNet are roughly consistent with learning rate differences in RB and II tasks. Each of these

measures is briefly described below. Readers not interested in the technical details can skip to the following section.

Difficulty measures

We know of only two previous articles that attempted to predict difficulty in II tasks (Alfonso-Reese, Ashby, & Brainard, 2002; Rosedahl & Ashby, 2019). Alfonso-Reese et al. (2002) focused on three measures (class separation, covariance complexity, and error rate of the ideal observer). Rosedahl and Ashby (2019) examined these same three measures, plus nine other measures that are popular in the machine-learning literature (described by Lorena, Garcia, Lehmann, Souto, & Ho, 2018), and in addition, they proposed a new measure (i.e., the striatal difficulty measure). Here we tested the predictions of all these measures on RB and II category structures. As mentioned earlier, we also tested AlexNet on these same RB and II categories (Krizhevsky et al., 2012). These 13 alternative measures and AlexNet are described briefly in this section. For more details, see Rosedahl and Ashby (2019), Lorena et al. (2018), and Krizhevsky et al. (2012).

Difficulty measures from the cognitive science literature

Class Separation (C_{sep} ; Alfonso-Reese et al., 2002) This measure is analogous to the between-category d' . It is defined as the distance between the category means divided by a measure of the standard deviation within the categories along this direction.

Covariance Complexity (CC; Alfonso-Reese et al., 2002) This is a measure that increases with the heterogeneity in the category variances and covariances.

Error Rate of the Ideal Observer (eIO; Alfonso-Reese et al., 2002) This is the error rate of a participant using the optimal classification strategy in the absence of any perceptual or criterial noise.

Striatal Difficulty Measure (SDM; Rosedahl & Ashby, 2019) SDM is defined as between-category similarity divided by within-category similarity, where between-category similarity is the summed similarity of all exemplars to all other exemplars belonging to contrasting categories and within-category similarity is the summed similarity of all exemplars to all other exemplars belonging to the same category.

Graph-theoretic measures of difficulty from the machine-learning literature

A number of difficulty measures that are popular in the machine-learning literature begin by representing the categories as a graph. In these approaches, each category exemplar is

represented as a node or vertex in a graph, and nodes are connected if the Gower distance in stimulus space between the corresponding exemplars is less than some criterion value. Finally, edges that connect exemplars from contrasting categories are pruned. In the current applications, Gower (1971) distance is similar to city-block distance that has been standardized to a [0,1] scale, with the value of 1 being assigned to the largest possible distance in the data set. So consider the Fig. 1 categories. Note that if the two categories are moved farther apart (in either condition), then the maximum possible distance will increase. Thus, the Gower distance between two exemplars from the same category decreases as the categories move farther apart.³ This means that, for the same pruning criterion, more exemplars will be connected to exemplars from the same category as category separation increases.

Average Density of the Network (Density; Lorena et al., 2018)

Density is defined as the number of edges in the graph divided by the maximum possible number of edges in a graph with the same number of nodes. With widely separated categories, little or no pruning will be necessary and density will be high. In contrast, with overlapping categories, many exemplars will have nearby neighbors that belong to the contrasting category, and therefore density will be low.

Clustering Coefficient (ClCoef; Lorena et al., 2018) This is a measure of average local density. First, for each node, define its neighborhood as the set of all nodes that are directly connected (i.e., in the same graph used to compute Density). ClCoef is the mean density of each of these neighborhoods.

Hub Score (Hubs; Lorena et al., 2018) The hub score equals the number of connections a node has weighted by the number of connections of each of its neighbors. With widely separated categories, each exemplar will be connected to more exemplars from its own category than with overlapping categories, so hub score increases with category separation.

Fraction of Borderline Points (FBP; Lorena et al., 2018) This is a related measure that uses a different algorithm to represent the categories as a graph or network. In this case, the categories are represented as a minimum spanning tree, which essentially is a tree constructed by placing an edge between each exemplar and its nearest neighbor, regardless of category membership. In this approach, there is no pruning. The fraction of borderline points equals the proportion of exemplars in all categories that are connected by an edge to an exemplar belonging to a contrasting category.

³ If two exemplars are in the same category then the absolute distance between them will not change as the categories are moved apart, but to standardize to a [0,1] scale we must divide their absolute distance apart by the maximum possible distance. Since this value increases with category separation, the Gower distance between the two exemplars must decrease.

Other machine-learning measures of difficulty

Collective Feature Efficiency (CFE; Orriols-Puig Macià, & Ho, 2010) This measure is based on the percentage of stimuli that can be correctly classified using decision bounds perpendicular to each stimulus dimension (i.e., the one-dimensional bounds that COVIS assumes are fundamental to the explicit rule-learning system).

Error Rate of Nearest Neighbor Classifier (eNN; Lorena et al., 2018) This is the error rate of a classifier that assigns the stimulus to the category of its nearest neighbor among all other stimuli in the two categories.

Fraction of Hyperspheres Covering Data (T_1 ; Ho & Basu, 2002) This measure is computed by first centering a hypersphere on each stimulus and setting the radius equal to the distance between that stimulus and the nearest exemplar from the contrasting category. All hyperspheres that are completely contained in another hypersphere are then removed and the measure is simply the fraction of original hyperspheres that remain.

Ratio of Intra- to Extra-Class Nearest-Neighbor Distance (N_2 ; Ho, 2002) N_2 equals the sum of the distances to the nearest neighbors in each contrasting category divided by the sum of the distances to the nearest neighbors in the same category.

Volume of Overlapping Regions (VOR; Souto, Lorena, Spolaôr, & Costa, 2010) This measure increases with the amount of overlap of the category distributions on each separate stimulus dimension.

AlexNet

AlexNet (Krizhevsky et al., 2012) is a deep convolutional neural network that includes five convolutional layers and three fully connected layers. We replaced the final three layers of the network with a fully connected layer, a softmax layer, and a classification output layer, respectively. To test the ability of AlexNet to learn the RB and II categories, we generated 600 stimuli from both structures, split the data sets into 70% (420 trials) training and 30% (180 trials) testing, and trained the network on ten minibatches of 42 trials each. We defined AlexNet difficulty as its error rate on the testing data.

Applications to RB and II Tasks

For AlexNet and each of these 13 measures, we computed difficulty of the RB and II tasks illustrated in Fig. 1. The results are shown in Table 1.⁴ Note that ten of the 14 measures

⁴ The numerical values in this table were computed from the RB and II categories shown in Fig. 1, which are identical to the categories used by Ell and Ashby (2006) in their medium-low condition.

Table 1 Predicted difficulty for the Fig. 1 rule-based (RB) and information-integration (II) categorization tasks

Difficulty Measure	RB Task	II task
AlexNet	0.005	0.002
Density	0.81	0.81
C_{sep}	0.07	0.07
ClsCoef	0.31	0.29
CFE	0	0.03
CC	0.32	0.32
eIO	0	0
eNN	0	0
FBP	0.003	0.003
T_1	0.01	0.01
Hubs	0.66	0.66
N_2	0.04	0.04
SDM	0.28	0.28
VOR	0	0.12

Note. Some of the measures we examined predict easiness, rather than difficulty (i.e., higher values indicate an easier task). Therefore, to aid interpretation, all such values in Table 1 are defined as 1 minus computed easiness. The one exception is C_{sep} , where we defined difficulty as the inverse of easiness. Thus, in all cases, higher values of all Table 1 measures indicate greater difficulty

predict that the RB and II tasks are equally difficult. It is not surprising that most of the measures predict no difficulty difference because the RB and II category structures are identical except for their orientation in stimulus space. The SDM is among these ten measures. This is noteworthy because the SDM is the best available predictor of learning difficulty in II tasks. As mentioned earlier, Rosedahl and Ashby (2019) showed that the SDM accounted for 87% of the variance in the final-block accuracies of 17 different data sets that were collected in different labs, used different stimuli, and included both linearly and nonlinearly separable categories. Unfortunately, this powerful measure of II task difficulty fails to account for the qualitatively strong finding that learning is much faster in the Fig. 1 RB task than in the II task, unless one acknowledges that participants bring qualitatively different learning processes to the RB task.

Supporting this idea, Rosedahl and Ashby (2019) showed that the SDM failed in one important instance – that is, human learners were better than the measure predicts on the Shepard, Hovland, and Jenkins (1961) Type 2 categories. Even Shepard and his colleagues were struck originally by the high levels of Type 2 performance. But it turns out that the optimal rule on Type 2 categories has a simple verbal description (e.g., respond A to a large square or a small triangle; otherwise respond B), so the Type 2 categories are best described as an RB task that is amenable to qualitatively different learning processes.

AlexNet actually performed slightly better on the II task. However, as mentioned previously, the AlexNet difficulty measure is its error rate in each task. Thus, the AlexNet accuracy was above 99% correct in both tasks, and as a result, the difference in predicted difficulty is negligible. Difficulty measure ClsCoef also predicts that the RB task is more difficult than the II task. ClsCoef is a network or graph-theoretic measure of task difficulty. As described in the previous section, the first step is to represent the categorization stimuli as a graph in which two stimuli are connected by an edge if their distance apart in stimulus space is less than some criterion value. The 45° rotation of the Fig. 1 RB categories that creates the II categories preserves the Euclidean distance between every pair of points, but not other types of distance. For example, the II category means are farther apart according to city-block distance than the RB means. The graph that ClsCoef operates on is built from a non-Euclidean metric that is similar to city-block distance (called Gower distance). This models the II categories to be more separated in perceptual space, and this is the reason that ClsCoef predicts that the RB task should be more difficult than the II task. Of course, this predicted difference is problematic because, empirically, RB tasks are generally more learnable, not less learnable.

Difficulty measures CFE and VOR predict that the II task is more difficult than the (rotated) RB task. Both of these measures privilege the stimulus dimensions over other directions through stimulus space. CFE predicts that difficulty decreases with the proportion of stimuli that can be correctly classified by a vertical or horizontal bound. Obviously, this proportion is 1 for the Fig. 1 RB categories, but less than 1 for the II categories. Similarly, VOR predicts that difficulty increases with the range of values on each dimension that are shared by both categories. Note that the RB categories share no common values on the bar-width dimension, whereas the II categories share common values on both dimensions. Thus, of the 13 difficulty measures considered here, only CFE and VOR correctly predict that with the Fig. 1 categories, RB learning will be faster than II learning. And both do so by acknowledging the learner's potential strong role in dimensional perception, selective attention, and rule learning.

The obvious next question, therefore, is whether either of these measures can accurately predict difficulty in all RB and II tasks. In other words, is either one of these measures a candidate for the unknown difficulty measure that must underly the difficulty hypothesis? The unfortunate answer is a definite no. Whereas the SDM accounted for 87% of the variance in the final-block accuracies of the 17 data sets examined by Rosedahl and Ashby (2019), VOR only accounted for 2% of this variance and CFE only accounted for 18%. Thus, neither measure that seems to accommodate the RB learning advantage is able to predict learning-rate differences across different II tasks. Both measures also make other incorrect predictions. For example, they also both incorrectly

predict no decrease in difficulty as the Fig. 1 RB categories are moved farther apart from each other.

In summary, the difficulty hypothesis requires as a prerequisite that some single quantitative measure of difficulty exists that simultaneously accounts for learning-rate differences in all RB and II tasks. If this is not true, then there is no basis to claim that the Fig. 1 RB task is less difficult than the Fig. 1 II task. In this section, we saw that no current measures of difficulty satisfy this prerequisite. The best available predictor of difficulty in RB tasks – namely the Boolean complexity of the optimal classification rule – is not even defined for II tasks. The best available predictor of difficulty in II tasks – the SDM – incorrectly predicts no difficulty difference between the Fig. 1 RB and II tasks. Furthermore, one popular difficulty measure actually predicts that the II task should be easier than the RB task, and nine other widely used measures predict no difference in RB and II difficulty. We did identify two measures that correctly predict that the II task is more difficult (CFE and VOR), but both measures failed miserably at predicting difficulty differences among different II tasks. In addition, we showed that one of the most powerful and popular deep convolutional neural networks also predicts no difficulty difference between the Fig. 1 RB and II tasks.

Does categorization difficulty depend on the state of the learner?

In the previous section, we saw that no current measures of categorization difficulty are able to account simultaneously for performance differences in RB and II tasks. Although this result supports the multiple-systems hypothesis over the difficulty hypothesis, it is not definitive because our results do not rule out the possibility that a single measure exists, but has not yet been identified. However, the second novel prediction derived in the section *Two novel predictions of the difficulty hypothesis* can be tested definitively. As we saw in that section, the difficulty hypothesis predicts that the ordering of tasks by performance can never reverse with changes in the state of the learner, whereas the multiple-systems hypothesis predicts that such reversals are possible.⁵

In fact, difficulty reversals have been documented in the literature. Many studies have shown that the Fig. 1 RB categories are learned much more quickly than the II categories – when the stimuli are the sine-wave gratings shown in Fig. 1. In fact, this large learning difference is the strongest evidence in support of the difficulty hypothesis. But it turns out that the Fig. 1 diagonal category structure is not always more difficult

⁵ The multiple-systems hypothesis predicts that difficulty must depend on the state of the learner, but it does not necessarily predict that these changes with the state of the learner have to be large enough to reverse some difficulty ordering.

than the rotated vertical (or, equivalently, horizontal) bound structures. Whether the Fig. 1 RB or II categories are more difficult to learn depends on how the participants perceive the stimulus dimensions.

For example, Ell, Ashby, and Hutchinson (2012) constructed two categories of color patches that varied on brightness and saturation. They started with a category structure that was almost identical to the Fig. 1 RB condition and then they created three other conditions by successively rotating the categories by 45°, 90°, and 135° in color space. In the original (i.e., 0° rotation) and 90° rotation conditions, only one stimulus dimension is relevant – either saturation or brightness. In contrast, both of these dimensions are relevant in the 45° and 135° conditions. The Ell et al. (2012) categories were essentially identical to those shown in Fig. 1. The only difference was in the name of the stimulus dimensions. Therefore, the difficulty hypothesis predicts that, as in Fig. 1, performance should be worse in the 45° and 135° conditions than in the 0° and 90° conditions. However, contrary to this prediction, the worst performance in the four conditions occurred in the 90° condition, in which the stimuli varied on only one relevant dimension (i.e., saturation).

Whether stimulus dimensions are perceptually separable or integral is a property of the perceiver, not of the stimuli themselves. Therefore, the Ell et al. (2012) results show that to predict whether the Fig. 1 RB or II categories will be more difficult requires knowledge about the state of the learner – namely, about how the learner perceives the stimulus dimensions. Thus, the Fig. 1 II task is not inherently more difficult than the RB task. The II task is more difficult for humans to learn with some stimuli, but not with others. COVIS accounts for this result by acknowledging that the nature of the stimulus dimensions profoundly affects the state of the learner, and a change in stimulus dimensions can change which learning system controls behavior. For the learner to apply the rule-learning system, the selected dimension must be perceptually separable from the non-selected dimension. For example, if we call the perceived value on the selected dimension Y , then the optimal strategy in the Fig. 1 RB categories is an explicit rule of the form “Respond A if $Y > y_c$, otherwise respond B,” where y_c is the criterion that separates the A and B categories. Note that this rule requires that the learner attend selectively to the attended dimension and that the perceived value on that dimension does not depend on the value of the stimulus on any non-selected dimensions. These are exactly the conditions required by perceptual separability (Ashby & Townsend, 1986). Saturation and brightness are integral dimensions. This makes attentional selection and rule-learning difficult or impossible. Thus, the Ell et al. (2012) results confirm that the perceptual-attentional state of the learner must also be considered when predicting task difficulty. This is strong evidence against the difficulty hypothesis and in support of the multiple-systems hypothesis.

No difficulty measure can account for dissociations between RB and II learning

We saw above that current measures of classification difficulty can predict learning rate differences in RB tasks, or in II tasks, but not concurrently in both, and next, we saw that whether the Fig. 1 RB task is easier or more difficult than the Fig. 1 II task depends on how participants perceive the stimulus dimensions. This section considers whether the difficulty hypothesis might account for the many RB and II dissociations if we simply accept that the Fig. 1 II task is more difficult than the Fig. 1 RB task. In other words, suppose that we accept that there is some measure of classification difficulty that correctly predicts difficulty in all RB and II tasks, but that this measure has not yet been discovered. If such a single difficulty measure existed, would the difficulty hypothesis be able to account for all the reported RB versus II dissociations?

The first problem with this counterfactual is the same one we had in the previous section – that is, that the Ell et al. (2012) results prove that such a measure cannot exist – at least, not if the difficulty measure is based only on statistical properties of the to-be-learned categories. This implies that the desired difficulty metric is unattainable. But in this section, we set this concern aside, and consider the validity of the difficulty hypothesis under the assumption that the categories are constructed from stimuli with perceptually separable stimulus dimensions. Thus, this section analyzes the literature in a manner strongly favorable to the difficulty hypothesis. Even so, the literature includes many results that are seemingly incompatible with any version of the difficulty hypothesis, and therefore there is strong evidence that the many reported dissociations between RB and II learning and performance are not due to an RB versus II difficulty difference.

As noted earlier, many studies have shown that under immediate feedback conditions, healthy young adults learn the Fig. 1 RB categories much more quickly than the II categories – at least when the stimuli are constructed from perceptually separable stimulus dimensions. If this RB over II advantage is due to a difficulty difference between the tasks, then any manipulation that impairs performance should have a greater effect on II learning than on RB learning. A manipulation that slows learning should impair a more difficult task more than an easy task. Many studies have disconfirmed this prediction of the difficulty hypothesis.⁶

First, if humans perform worse in the II task because it is more difficult or complex than the RB task, then the RB versus II difference should be even greater in a cognitively simpler species. Tasks that are difficult for humans should be even

⁶ This prediction of the difficulty hypothesis is not unique to this article (e.g., Smith et al., 2011), which is why it is not included in the section *Two novel predictions of the difficulty hypothesis* as a third novel prediction. Nevertheless, to our knowledge, this is the first attempt to test this prediction against a wide variety of different data.

more difficult for a simpler species. In contrast to this prediction, two different labs independently showed that pigeons learn rotated RB and II categories at exactly the same rate (Smith et al., 2011). Recently, a third study replicated this finding (Qadri, Ashby, Smith, & Cook, 2019). This striking identity of learning performance is shown in Fig. 2. Thus, the performance of pigeons is perfectly predicted by the SDM and many of the other Table 1 difficulty measures. The orientation of the decision bound is irrelevant to pigeons. Of course, pigeons may use a different strategy in both tasks than humans, but their identity of performance supports the hypothesis that they used the same strategy in both tasks, and that as predicted by most difficulty measures, the two tasks are equally difficult.

COVIS predicts that humans are better in the RB task because they use their well-developed rule-learning system. This system fails in the II task, so humans must resort to a slower, associative-learning mechanism. Presumably, pigeons lack an effective rule-learning system, and therefore must resort to some form of associative learning in both tasks. Evidence supporting this comes from Berg, Ward, Dai, Arantes, and Grace (2014) who compared pigeon and human performance in the Fig. 1 RB task and its 90° rotation, in which the single relevant dimension was bar orientation. They reported that “Results showed that humans learned both tasks faster than pigeons, with abrupt increases in accuracy that were indicative of rule-based responding, while pigeons learned the tasks gradually” (p. 44). Another possibility, however, is that pigeons perceive the Gabor stimuli in a fundamentally different way than humans. Perhaps for pigeons, bar width and orientation are not fundamental perceptual dimensions and therefore, for pigeons, both Fig. 1 tasks require attention to two perceptual dimensions. Qadri et al. (2019) reported evidence against this hypothesis. They reported that pigeons showed greater perceptual generalization along the dimensions of bar width and

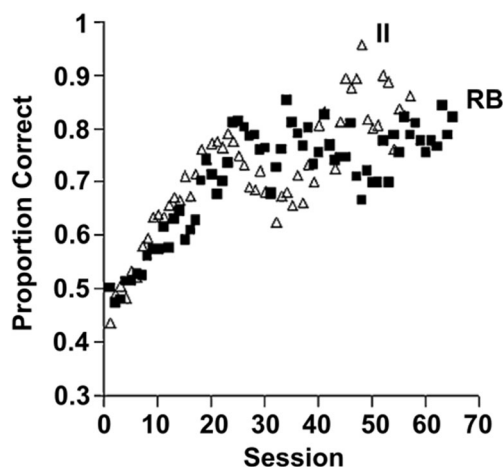


Fig. 2 Pigeons' learning curves in information-integration (II) and rule-based (RB) categorization tasks. The proportion of correct responses in each session are shown from the criterial block backward for eight II-learning pigeons and eight RB-learning pigeons. Reprinted from Smith et al. (2012)

orientation than along other directions, and they concluded that “these dimensions are salient, independent, and meaningful in the pigeons' categorization” (p. 265).

Furthermore, Broschard, Kim, Love, Wasserman, and Freeman (2019) recently reported similar results with rats. Specifically, they reported that rats also learn RB and II categories like those shown in Fig. 1 equally well and at exactly the same rate. This is an important result because rats are mammals and therefore have a brain structure that is more similar to humans than pigeons. Broschard et al. (2019) concluded that “rats extract and selectively attend to category-relevant information but do not consistently use rules to solve the RB task” (p. 84). Thus, rats, like pigeons, seem to learn RB and II categories in the same way. In summary, neither pigeons nor rats find the II task to be more difficult than the RB task, which strongly suggests that there is no inherent difficulty or complexity difference between the tasks, and therefore that the human RB advantage occurs, not because of a difficulty difference, but because people learn the two tasks in qualitatively different ways.

Second, several studies have reported that a simultaneous (working-memory) dual task interferes more with one-dimensional RB learning than with II learning (e.g., Crossley et al., 2016; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). This is opposite to what the difficulty hypothesis predicts. Trying to do two things at once should be harder if the two tasks are difficult than if they are easy. And the fact that the second task in these studies required working memory is especially problematic for the difficulty hypothesis because the presumed difficulty difference between RB and II tasks is often justified by the assumption that the II task has greater working memory demands than the RB task. For example, Le Pelley et al. (2019) cited the “low memory-demands associated with the” RB category-learning task as a major driver of the RB versus II difficulty difference. If the II task is more difficult than the RB task because it has higher memory demands, then increasing working memory load via a simultaneous dual task should have a much more deleterious effect on the more memory-demanding II task. But the opposite is clearly true. A simultaneous (working-memory-dependent) dual task interferes with learning much more in the “simple” RB task than in the “difficult” II task. It is important to note that COVIS predicts these results *a priori*. According to COVIS – the most influential current instantiation of the multiple-systems hypothesis – learning the optimal rule in the RB task requires working memory and executive attention, whereas executive function plays little or no role in the procedural learning needed to optimize performance in the II task.

Third, a number of studies have shown that manipulations that interfere with feedback processing impair RB learning more than II learning (Maddox, Ashby, Ing, & Pickering, 2004a; Zeithamova & Maddox, 2007). This result also runs counter to the difficulty hypothesis. The II task, the putatively more difficult task, should need more of the effortful and

attentional feedback processing than the RB task, so the II task should suffer more when those interpretative processes are blocked. The opposite is clearly true. COVIS correctly predicts these results (again in an *a priori* fashion) because it assumes that II learning is about the building of associative connections between cortical stimulus representations and behavioral outputs. Feedback is incorporated automatically into strengthening these associations (Maddox et al., 2003). Blocking post-trial cognition will not affect this process. In contrast, humans in the RB task are testing, (dis)confirming, and revising dimensional hypotheses to learn the appropriate rule. They must effortfully evaluate what feedback means for their ongoing hypothesizing, and so derailing post-trial cognition has serious consequences.

Fourth, pressure to perform impairs performance in RB tasks, and actually improves performance in II tasks (Markman, Maddox, & Worthy, 2006; Worthy, Markman, & Maddox, 2009). Similarly, Ell, Cosley, and McCoy (2011) reported that stress improved performance and the use of optimal strategies in II tasks, but impaired performance in RB tasks, although not significantly. These results are strongly incompatible with the difficulty hypothesis. For example, the difficulty hypothesis predicts that the difficult II task should crumble first under pressure, yet the opposite is true. Instead, these results are compatible with multiple learning systems. The implicit system learns automatically – pressure cannot affect the forming of stimulus-response associations. Moreover, under pressure, one is less likely to send the explicit learning system to hijack the task, which can lead to adventitious rules and poor performance. So, true II learning can occur. However, pressure might well derail the explicit, effortful cognitive processes that are necessary to solve the RB task, and so it could be impaired.

Fifth, increasing the number of contrasting categories slows RB learning but not II learning (Maddox, Filoteo, Hejl, & David, 2004b) – again, exactly opposite to the prediction of the difficulty hypothesis. Sixth, many studies have deliberately included more difficult RB conditions – so that RB and II performance is approximately equated – and found that the qualitative differences predicted by COVIS are still evident (e.g., Crossley et al., 2016; Maddox & Ing, 2005).

Conclusions

Difficulty is a tricky construct. It seems so familiar that one is tempted to apply the criterion that Supreme Court Justice Potter Stewart used to identify obscenity in *Jacobellis v. Ohio* – namely, “I’ll know it when I see it” (Lattman, 2007). Examining Fig. 1 can produce the impression that the II task *is* more difficult than the RB task. After all, learning the RB categories is simply a matter of figuring out that Category A disks have thick bars and Category B disks have thin bars, whereas learning the II categories requires ... well, something

else. But this is exactly the point of the multiple-systems hypothesis. Learning the RB categories is a matter of finding the optimal explicit rule, whereas no such rule exists for the II categories. With RB categories, humans can accelerate learning by exploiting our sophisticated logical reasoning abilities. The RB task seems easier because we evolved a system that is perfectly suited for this task. With the II categories, no logical rule determines category membership so this system fails. Instead, II categories must be learned in a slower, more incremental and associative fashion.

The “I’ll know it when I see it” criterion is dangerous because it flirts with circularity. One can always examine any set of results, note which condition has the worst performance, and then assert by fiat that this is the most difficult condition. But this approach has no predictive power – the results are used as an explanation of themselves.

Difficulty is a useful scientific construct only if it can be defined rigorously. Unfortunately, none of the studies hypothesizing that the RB versus II dissociations are due to difficulty differences defined what they mean by difficulty. The vague theoretical nature of this claim makes it much more difficult to test. Nevertheless, as demonstrated here, strong tests are possible. And, as we have shown, these tests overwhelmingly reject the difficulty hypothesis and instead strongly favor the multiple-systems account of the many RB versus II dissociations. Even so, it is important to acknowledge that any single article is unlikely to definitively disconfirm or confirm any psychological theory, and in this sense, the ambiguity of the difficulty hypothesis works in its favor. For example, we showed that a wide variety of current difficulty measures fail to account for learning rate differences between RB and II tasks, but we did not show that a difficulty measure that can account for these differences does not exist. And we disconfirmed several predictions that seem to follow in a straightforward manner from the difficulty hypothesis, but perhaps some other subtler version of the hypothesis is possible that is not constrained by these predictions. Even so, our results make an important contribution because they show that the most straightforward interpretation of the difficulty hypothesis is incompatible with many well-established empirical phenomena. In this sense, the ball is now in the court of proponents of the difficulty hypothesis. We found no evidence supporting this hypothesis and we disconfirmed several of its seemingly straightforward predictions. Resurrecting this theoretical account of the many reported dissociations between RB and II tasks now requires a more formal description of the difficulty hypothesis that overcomes the apparent weaknesses that were demonstrated here.

Author Notes Preparation of this article was supported by the National Institutes of Health grant 2R01MH063760 to FGA, the NICHD grant R01-HD093690 to JDS, and the National Science Foundation Graduate Research Fellowship under grant number 1650114 to LAR.

References

- Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, *64*, 570–583.
- Ashby, F. G., Alfonso-Reese, L. A., Turken, A., & 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., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological review*, *93*, 154–179.
- Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen and C. Lefebvre (Eds.), *Handbook of categorization in cognitive science, 2nd Edition* (pp. 157–188). Cambridge, MA: Elsevier.
- Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*, 363–378.
- Berg, M. E., Ward, M. D., Dai, Z., Arantes, J., & Grace, R. C. (2014). Comparing performance of humans and pigeons in rule-based visual categorization tasks. *Learning and Motivation*, *45*, 44–58.
- Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019). Selective attention in rat visual category learning. *Learning & Memory*, *26*, 84–92.
- Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1388–1403.
- Crossley, M. J., Paul, E. J., Roeder, J., & Ashby, F. G. (2016). Declarative strategies persist under increased cognitive load. *Psychonomic Bulletin & Review*, *23*, 213–222.
- 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.
- 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.
- 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.
- 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.
- 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*, 96–102.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–871.
- Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*, 1013–1031.
- Ho, T. K. (2002). A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, *5*, 102–112.
- Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 289–300.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- Lattman, P. (September 27, 2007). The origins of Justice Stewart’s ‘I know it when I see it’. *Wall Street Journal. LawBlog at The Wall Street Journal Online*.
- Le Pelley, M. E., Newell, B. R., & Nosofsky, R. M. (2019). Deferred feedback does not dissociate implicit and explicit category learning systems: Commentary on Smith et al. (2014). *Psychological Science*, in press.
- Lorena, A., Garcia, L., Lehmann, J., Souto, M., & Ho, T. (2018). How complex is your classification problem? A survey on measuring classification complexity. *arXiv*.
- 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.
- Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004a). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, *32*, 582–591.
- Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004b). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 227–245.
- 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.
- Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science*, *17*, 944–948.
- 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.
- Orriols-Puig, A., Macià, N., & Ho, T. K. (2010). Documentation for the data complexity library in C++. Technical report, La Salle - Universitat Ramon Llull.
- Qadri, M. A. J., Ashby, F. G., Smith, J. D., & Cook, R. G. (2019). Testing analogical rule transfer in pigeons (*Columba livia*). *Cognition*, *183*, 256–268.
- Rosedahl, L., & Ashby, F. G. (2019). A difficulty predictor for perceptual category learning. *Journal of Vision*, *19*(6), 20.
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.
- 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., Berg, M. E., Cook, R. G., Boomer, J., Crossley, M. J., Murphy, M. S., Spiering, B., Beran, M. J., Church, B. A., Ashby, F. G., & Grace, R. C. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience and Biobehavioral Reviews*, *36*, 2355–2369.
- Smith, J. D., Boomer, J., Zakrzewski, A., Roeder, J., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447–457.
- Souto, M. C. P., Lorena, A. C., Spolaôr, N., & Costa, I. G. (2010). Complexity measures of supervised classification tasks: A case study for cancer gene expression data. In International Joint Conference on Neural Networks (IJCNN), 1352–1358.
- 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.
- Worthy, D. A., Markman, A. B., & Maddox, W. T. (2009). What is pressure? Evidence for social pressure as a type of regulatory focus. *Psychonomic Bulletin & Review*, *16*, 344–349.

- Zaki, S. R., & Kleinschmidt, D. F. (2014). 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.
- Zeithamova, D., & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition, 35*, 1380–1398.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.