Journal of Experimental Psychology: Learning, Memory, and Cognition

Identifying Strategy Use in Category Learning Tasks: A Case for More Diagnostic Data and Models

Chris Donkin, Ben R. Newell, Mike Kalish, John C. Dunn, and Robert M. Nosofsky

Online First Publication, December 22, 2014. http://dx.doi.org/10.1037/xlm0000083

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2014 American Psychological Association
0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000083

Identifying Strategy Use in Category Learning Tasks: A Case for More Diagnostic Data and Models

Chris Donkin and Ben R. Newell
University of New South Wales

Mike Kalish
Syracuse University

John C. Dunn
University of Adelaide

Robert M. Nosofsky
Indiana University

The strength of conclusions about the adoption of different categorization strategies—and their implications for theories about the cognitive and neural bases of category learning—depend heavily on the techniques for identifying strategy use. We examine performance in an often-used "information-integration" category structure and demonstrate that strategy identification is affected markedly by the range of models under consideration, the type of data collected, and model-selection techniques. We use a set of 27 potential models that represent alternative rule-based and information-integration categorization strategies. Our experimental paradigm includes the presentation of nonreinforced transfer stimuli that improve one's ability to discriminate among the predictions of alternative models. Our model-selection techniques incorporate uncertainty in the identification of individuals as either rule-based or information-integration strategy users. Based on this analysis we identify 48% of participants as unequivocally using an information-integration strategy. However, adopting the standard practice of using a restricted set of models, restricted data, and ignoring the degree of support for a particular strategy, we would typically conclude that 89% of participants used an information-integration strategy. We discuss the implications of potentially erroneous strategy identification for the security of conclusions about the categorization capabilities of various participant and patient groups.

Keywords: category learning, categorization, model selection

Perceptual category learning refers to the process by which we learn to organize different perceptual objects into distinct groups. The cognitive machinery underpinning this ability has been the focus of extensive research, and a large part of this effort has involved attempts to develop computational models that can account for our capacity to categorize (Pothos & Wills, 2011). Often, alternative models are fitted to classification data in an attempt to identify the categorization strategy used across different conditions of testing and by different populations of subjects, such as younger versus older adults, or patients with different neurological disorders. We argue here that such identification would benefit from improvements in the type of data used to fit models, the procedures for fitting and selecting among those models, and the range of models under consideration. This reevaluation of data, models and methods has significant implications for theories of category learning and for the security of conclusions about the categorization capabilities of various participant groups.

We illustrate and demonstrate the value of our approach by examining some of the evidence presented in support of one the most influential models of perceptual categorization: COVIS (Competition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998). According to COVIS, categories can be acquired via a verbal system that generates and tests simple verbalizable hypotheses, or rules, and depends on structures in the anterior cingulate, the prefrontal cortices, the medial temporal lobe, and the head of the caudate nucleus (Ashby & Ell, 2001; Ashby & Spiering, 2004; Nomura & Reber, 2008). In addition, categories can be learned via a procedural system that learns to associate a response with regions of perceptual space based on reinforcement (Ashby, Paul, & Maddox, 2011) and depends on neural structures in the tail of the caudate nucleus (Ashby et al., 1998; Nomura & Reber, 2008).

Two types of category tasks are typically used to index the hypothesized differential engagement of these systems in category learning. In "rule-based" (RB) tasks, optimal performance can be achieved via the use of sets of logical rules for partitioning the objects into categories. In implementing such logical rules, the observer makes separate decisions about the values of objects along their component dimensions and then combines those decisions to determine which rule has been satisfied. For example, lines varying in their length and orientation might be assigned to

Chris Donkin and Ben R. Newell, School of Psychology, University of New South Wales; Mike Kalish, Department of Psychology, Syracuse University; John C. Dunn, School of Psychology, University of Adelaide; Robert M. Nosofsky, Department of Psychological & Brain Sciences, Indiana University.

Correspondence concerning this article should be addressed to Chris Donkin, School of Psychology, Matthews Building, UNSW, Kensington, 2052 Australia. E-mail: christopher.donkin@gmail.com

Category A only if they are sufficiently long and sufficiently steep (i.e., a conjunctive rule). We provide examples of a wide variety of such RB strategies in the modeling-analysis section of our article. Because such rules are generally readily verbalizable, COVIS theorists hypothesize that these RB strategies are learned by the verbal categorization system.

In contrast, in "information-integration" (II) tasks, optimal performance requires that information from more than one dimension be perceptually integrated before any categorization decisions are made (Ashby et al., 1998). For example, one might compare the overall similarity of an object to the prototypes of alternative categories and classify the object into the category with the nearest prototype. Again, there is a wide variety of potential II strategies and we provide numerous examples in the modeling-analysis section. Because the different II strategies are often difficult to verbalize, COVIS theorists hypothesize that II strategies are learned via the procedural system.

There is an extensive literature that has demonstrated a variety of experimental dissociations between performance on RB and II tasks. These dissociations have been taken as evidence in support of the idea that RB and II tasks are generally learned by the separate systems hypothesized in the COVIS theory. However, the interpretation of these dissociations has been the subject of considerable recent debate (e.g., Dunn, Newell, & Kalish, 2012; Newell, Dunn, & Kalish, 2010, 2011; Nosofsky, Stanton, & Zaki, 2005; Zaki & Kleinschmidt, 2014). This debate is important for a number of reasons, but not least because in recent years II and RB tasks have been used as "diagnostic tools" to examine how category learning is affected by various neurological conditions (e.g., Ell, Marchant, & Ivry, 2006; Huang-Pollock, Maddox, & Tam, 2014; Maddox et al., 2011; Schnyer et al., 2009). Many of these studies are inspired by the neurobiological bases of the COVIS model and seek to find evidence that damage to brain regions hypothesized to underpin the verbal and procedural systems lead to characteristic detriments in category learning. For example, Ell et al. (2006) predicted and found that patients with a focal lesion to the basal ganglia due to stroke were impaired relative to controls in learning an RB task but that both groups performed equally on an II task.

However, as Ell et al. (2006) and other authors (Maddox et al., 2011; Schnyer et al., 2009) emphasize, raw performance accuracy (e.g., proportion correct) is but one, relatively crude measure of how participants might differ in learning these tasks. A much richer assay of performance can be gleaned via the application of model-based strategy analyses. As stated by Maddox, Pacheco, Reeves, Zhu, and Schnyer, (2010, p.2999): "models provide important insights onto the cognitive processes and strategies being utilized by participants to solve each task. Importantly, this information cannot be garnered from an examination of performance accuracy alone, as qualitatively different strategies often yield the same performance level."

In most cases, the standard practice for model fitting in these studies involves comparing the fit of a single II model to the fits of two types of RB models to data from II and RB tasks. To illustrate this practice, consider the stimulus structure shown in the left panel of Figure 1. This category structure, introduced by Maddox, Filoteo, Hejl, and Ing (2004), provides the stimuli for an information-integration task with four separate regions of perceptual space, each constituting a different category (A, B, C, & D). The stimuli themselves are lines that differ in length and orientation. The rule-based version of this category structure is a rotated version of this space wherein the category boundaries are vertical and horizontal rather than diagonal (see Maddox et al., 2004). These structures have been used in numerous recent studies, including the Ell et al. (2006) one described above, as well as the Maddox et al. (2010) study of age-related decline in category learning, and the Schnyer et al. (2009) study of patients with damage to the ventral prefrontal cortex.

In the Schnyer et al. (2009) study the authors found that although there was an overall accuracy detriment when comparing

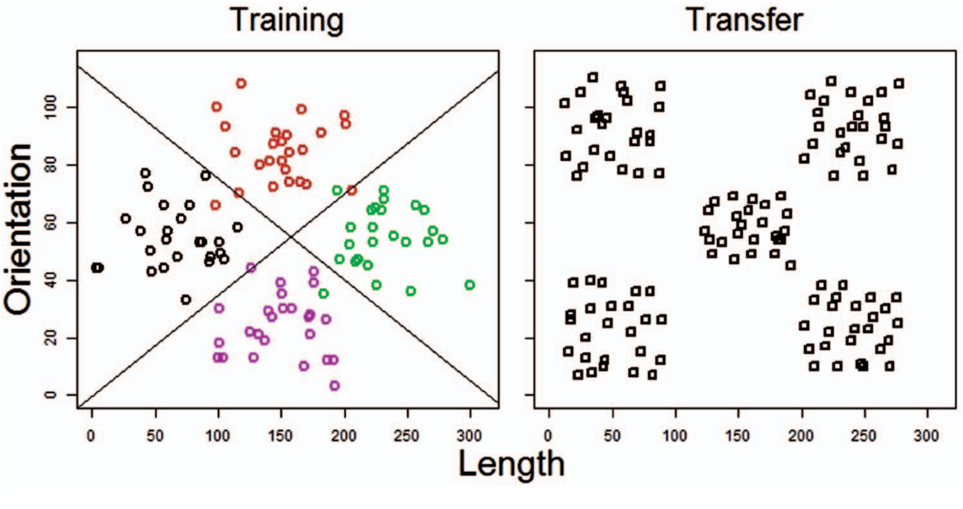

Figure 1. The training (left) and transfer (right) stimuli used in our experiment plotted as a function of the length and orientation of each line. The colors used for training stimuli indicate the category to which they belong (A = black, B = red, C = green, D = violet). The transfer stimuli belonged to no category. The diagonal lines drawn with the training stimuli reflect approximately optimal II classification boundaries. See the online article for the color version of this figure.

patients to controls on both II and RB tasks, model-based analyses revealed that better performance of control participants was due to controls adopting the optimal strategy in a task (e.g., using an II strategy in an II task). Similarly, Maddox et al. (2010) used model-based analyses to argue that age-related deficits on an II task were due to younger adults more consistently applying an II strategy compared to older adults. Likewise, Ell et al. (2006) concluded that the impairments in learning an RB task—especially in the early trials—shown by the basal ganglia patients was due to control participants showing more optimal use of RB strategies in RB tasks.

Naturally, the strength of conclusions about the adoption of different strategies—and their implications for theories about the cognitive and neural bases of category learning—depend heavily on the techniques used to identify strategy adoption. For example, in the studies just outlined, many of the conclusions rely on control participants being reliably identified as using the appropriate II strategy in an II task. Our contention is that existing studies fall short in exploring the range of possible strategies that participants could adopt in these tasks. If our contention is correct, the concern raises potential questions about the security of various conclusions regarding the categorization abilities of different groups of participants.

We focus our experimental and modeling analysis on the II structure presented in the left panel of Figure 1. We chose this structure because of its prevalence in recent studies (as discussed) and because it affords a wide range of potential strategies. Although an II strategy is optimal, a variety of alternative RB strategies could yield similar levels of performance. Failure to consider this full spectrum of RB strategies may lead to overestimates of the number of participants adopting II strategies.

We see two reasons why such incorrect identification of strategies may have occurred in previous studies using this and similar II structures. First, in the standard experiments using these tasks participants are given a series of training trials in which stimuli are repeated across blocks (e.g., six blocks of 100 trials in which each of the 100 stimuli are presented once per block). While this procedure provides a good index of learning, the absence of transfer trials, in which novel and—crucially—nonreinforced stimuli are presented limits conclusions about the types of strategies that participants adopt. For example, a participant given the structure in the left panel of Figure 1 might generate four orthogonal boundaries. Each category would correspond to an extreme value along each dimension, marked by those boundaries (i.e., far to the left, far to the right, far up, or far down). The application of such a strategy would leave relatively few stimuli for which responding would be ambiguous (i.e., different rules point to different responses). The limitation of the standard designs is that there are too few data points to sharply discriminate between such a RB strategy and the assumed II strategy. That is, the participant is almost never asked to classify a stimulus for which the II and RB models make sharply divergent predictions.

Our solution to this problem is to introduce a block of transfer trials following standard training (cf. McKinley & Nosofsky, 1996). The right panel of Figure 1 shows the areas of the stimulus space from which these transfer stimuli are drawn. As can be seen, the stimuli occupy those regions of the space that are underrepresented (or absent) in the training trials. As we soon demonstrate, these transfer stimuli occupy regions of the stimulus space that sharply contrast the potential RB and II strategies for classifying the training stimuli. As explained in detail in the Method and Procedure sections, we withheld corrective feedback on these transfer trials to prevent any learning about these stimuli, thereby gaining a pure measure of the boundaries participants adopt as a result of training.

The second reason for possible misidentification of strategy use is simply the limited number of potential models fit to the data in previous experiments. Many papers (e.g., Schnyer et al., 2009) compare the fits of only two variants of an RB model and one version of an II model. We include these models in our own analyses but embed them in a much wider range of both RB and II variants. We investigate this extended set of models in recognition of the fact that there are many ways a participant can represent the stimulus space depicted in Figure 1 and still achieve satisfactory performance. Finally, as will be seen, in drawing conclusions about strategy use, our model-selection methods will incorporate uncertainty in evaluating the extent to which competing models account for the data.

In summary, we conducted an experiment in which participants were given stimuli generated by the II structure shown in Figure 1. They learned to classify these stimuli across 400 training trials and then completed a further 249 transfer trials. We then fit a variety of RB and II models to individual data profiles. Our aim was to gain greater insight than has previously been provided into the ways participants solve this information-integration task.

Method

Participants

Sixty-two first-year psychology students from UNSW participated in return for course credit. (We chose to collect data for 4 days and the 62 subjects were those who agreed to participate during this time span.)

Design and Stimuli

Participants classified black lines of varying length and orientation, presented on a white background, into four categories. In the first phase of the experiment, participants completed four blocks of 100 training trials. Each of the 100 stimuli shown in the left panel of Figure 1 was presented once per block during training. The lengths and orientations of the lines were taken directly from the bottom right panel of Figure 1 in Schnyer et al. (2009). The color used to represent each stimulus in the left panel of Figure 1 reflects the category to which the line belongs (category A stimuli are shown in black, category B in red, C in green, and D in violet). In the second phase of the experiment, participants completed a further 249 test trials. Of the test trials, half (124) were the familiar training stimuli, with each of the 100 training stimuli presented once and a random sample of 24 training stimuli (without replacement) presented for a second time. On the remaining 125 trials, participants were presented with the transfer stimuli shown in the right panel of Figure 1. The transfer stimuli did not belong to any category, and so are all plotted using a square symbol.

The transfer stimuli were chosen to help identify the strategy participants were using to classify stimuli during the first phase of the experiment. In particular, we first outlined rectangles in the four corners and in the center of the stimulus space, the dimensions of which are given in Table 1 (see also the right panel of Figure 1). These large rectangles were then divided evenly into 25 smaller rectangles, and a random combination of length and orientation was sampled from within each small rectangle to create a single transfer stimulus. The same set of transfer (and training) stimuli was presented to all participants, and these are shown in Figure 1.

Procedure

Trials began with the presentation of the stimulus in the center of the screen, where it remained until a response was made. The participant was asked to respond as to which category, A, B, C, or D, they believed the stimulus belonged using the response keys Q, P, D, or K, respectively. After the response, if the participant was presented with one of the training stimuli then they were told whether their response was correct or incorrect, and the correct category was presented in the center of the screen for 1 second. If the item was a transfer stimulus, then the participant was simply given the feedback "Okay" for 1 second. After feedback the screen turned blank for 500 ms before the next trial began.

At the beginning of the experiment, participants were instructed that they would be classifying lines into four categories. They were told to place their index fingers on the D and K keys and middle fingers on Q and P. Participants were given further instruction at the beginning of each of the two test blocks, and were told that on the upcoming trials there would be some trials on which they would only receive the feedback "Okay," but that they would continue to receive the usual feedback on other trials.

Results

Trials with response times shorter than 200 ms or longer than 5,000 ms were excluded from analysis for being either unrealistically fast or slow, respectively, which led to 2.85% of the data being censored.

As shown in Figure 2, the proportion of correct responses changed across the blocks of the experiment ($BF = 1 \times 10^{56}$), initially increasing across the four training blocks from 0.60 to 0.83 ($BF = 7 \times 10^{20}$), but then decreased to 0.78 ($BF = 2.9 \times 10^{5}$) in the final test block.[1]

Figure 3 shows the proportion of responses made for each category for training stimuli (left panel), transfer stimuli (center panel), and for all stimuli simultaneously (right panel). Each location in the plot represents an individual stimulus. For each stimulus, there are up to four colored circles. The color of each circle represents the category response given to that stimulus. The size of each circle represents the proportion of times that a particular response was chosen, aggregated across participants and trials. For example, the stimuli with the shortest length were almost unanimously classified into category A, and so the circles for those stimuli are large and black. The stimuli toward the center of the stimulus space, with moderate orientation and length, were classified less consistently, and so there are up to four smaller circles of different colors.

Table 1
Coordinates in Stimulus Space Used to Construct Transfer Stimuli

Location in stimulus space	Dimension			
	Length$_1$	Length$_2$	Orientation$_1$	Orientation$_2$
Bottom left	10	90	5	40
Top left	10	90	75	110
Center	120	200	45	70
Bottom right	200	280	5	40
Top right	200	280	75	110

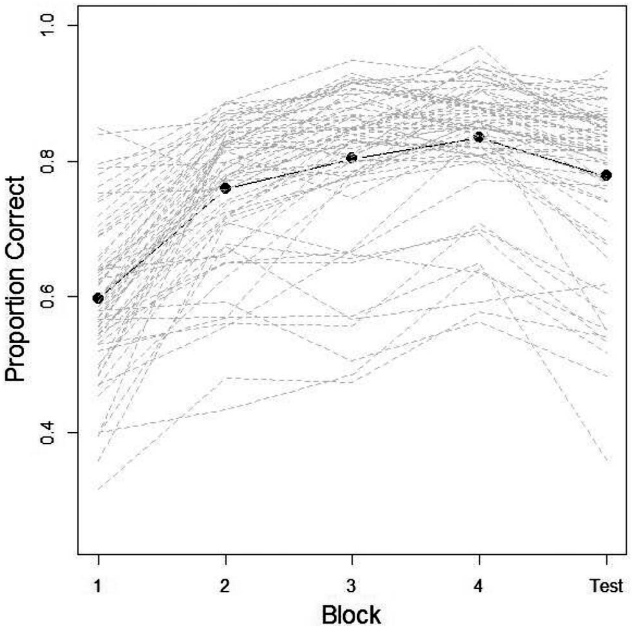

Figure 2. Proportion correct responses for training stimuli as a function of block number. Blocks 1 to 4 were training blocks, while the fifth block was a test block containing these training stimuli and additional transfer stimuli. The average of all participants is plotted in black and individual participant performance is plotted in gray.

Looking at the left panel of Figure 3, we see that on average participants were relatively accurate, particularly for stimuli that were further from the center of the stimulus space (i.e., the circles are large and of the appropriate color). On the other hand, responses to the transfer stimuli were more variable. There are relatively few transfer stimuli that were consistently classified into a single category by all participants. The right panel of the figure, which plots all stimuli together, suggests that the diagonal boundaries of an II classification strategy may have been utilized. However, this result does not necessarily imply that all participants used this strategy, as the figure plots the average classification of all participants. We now turn to a model-based analysis of our data to investigate the classification strategies of individuals.

Model-Based Analysis

We fit a range of variants of both II and RB models to the data from our experiment. The RB model variants were chosen to

[1] Bayes Factors were calculated using Richard Morey's BayesFactor package, using JZS default priors (see Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder, Morey, Speckman & Province, 2012).

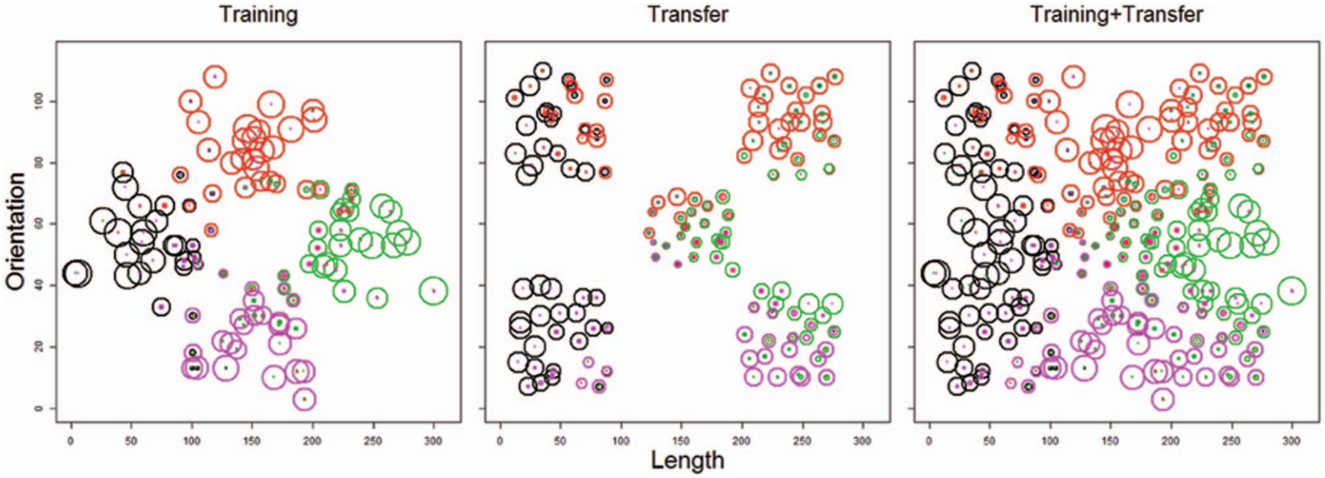

Figure 3. Proportion of responses for each category (colors: A = black, B = red, C = green, D = violet) plotted as circles of different sizes for each of the training (left panel), transfer (center), and all stimuli (right). See the online article for the color version of this figure.

represent the various rules by which participants may have classified the stimuli from our experiment, and we describe each of these below. We also fit a number of II model variants so as to avoid having to commit to one of the many possible II strategies that might be employed.

Rule-based models. All RB models assume that stimuli are represented using bivariate normal distributions, with mean equal to the stimulus location, and standard deviation determined by free parameters s_X and s_Y. On any given trial, the stimulus is perceived to be in the location that corresponds to a random sample from this normal distribution. A rule is implemented by assuming that vertical and horizontal response boundaries divide the stimulus space into regions that correspond to different category responses. Then, on any given trial, whichever region the stimulus appears to be in on that trial determines the response the participant makes.

So, for example, one rule may dictate that any stimulus with length smaller than X_1 is a member of category A, regardless of the orientation Y. Now consider a stimulus with true location ($x = a$, $y = b$). The probability that the participant makes a category A response to that stimulus is determined by the area under a bivariate normal distribution with mean (a,b) and standard deviation (s_X,s_Y) in the region $x < X_1$.

In the first set of RB models (RB1), we assumed that the participant would have a simple rule for classifying extreme stimuli on one dimension, and a more complex rule for intermediate stimuli. In RB Model 1a, we assumed that any stimulus[2] with length shorter than X_1 or longer than X_2 would be classified as belonging to either category A or C, respectively. Any stimulus with length between X_1 and X_2 was given the category B or D label, depending on whether the orientation of the line was steeper than Y_2 or flatter than Y_1, respectively. Any stimulus that falls within the center of the stimulus space (i.e., between both X_1 and X_2, and Y_1 and Y_2) is randomly assigned a category label. In RB Model 1b, we assumed that now the orientation dimension was given priority, and that any stimulus with orientation steeper than Y_2 or flatter than Y_1 would be classified into category B or D, respectively. Intermediate stimuli are placed into categories A or C depending on whether they are shorter than X_1 or longer than X_2, respectively. Finally, in RB Model 1c, we assumed that participants would use a mixture of the strategies applied in Models 1a and 1b. On half of trials, participants would prioritize the length dimension, and on the other half of trials, participants prioritize orientation. This could be considered as a mixture of two rules.

In the second set of RB models (RB2), we assume that participants do not prioritize any given dimension, but instead divide the space so that any stimulus with length shorter than X_1 or longer than X_2 would be classified as A or C, and any stimulus with orientation steeper than Y_2 or flatter than Y_1 would be classified as B or D. However, this leaves four regions of space that lead to contradictory classifications. For example, any stimulus shorter than X_1 and steeper than Y_2 should be called either A or B. We assume that participants guess between these two options whenever a stimulus falls into this region of the space. Again, we assume that any stimulus that falls into the center of the space is randomly assigned one of the four category labels.

In a third set of RB models (RB3), we assumed that one dimension was given priority, and hence a simple rule, and that the remainder of the space was divided into the three remaining categories. So, in RB Model 3a, we assume that any stimulus shorter than X (since $X_1 = X_2$ in this set of models) is given the label A. Any stimulus longer than X is called B if it has orientation steeper than Y_2, C if it has orientation between Y_1 and Y_2, and D if its orientation is flatter than Y_1. There were four versions of RB3: 3a corresponds to the simple rule being applied to stimuli being shorter than X, 3b to stimuli being steeper than Y, 3c to stimuli being longer than X, and 3d to stimuli being flatter than Y.

The RB Models 1a–1c and 2 require that the participant sometimes guess between two or more responses. We also fit a second version of these models in which we assumed that participants might be biased to respond with certain category labels over

[2] Throughout this section, the term *stimulus* denotes the perceived stimulus, not the objective one.

others. To incorporate this idea, we estimated three additional parameters – b_{A}, b_{B}, and b_{C}, setting b_{D} to be 1 (without loss of generality). The overall level of bias, B_i, toward category i was given by $B_i = \frac{b_i}{b_A+b_b+b_c+1}$. Then, whenever a guess was required between categories j = A, B . . . the bias toward a particular category was given by $\frac{B_i}{\Sigma_j B_j}$. So, for example, if a guess was required between category A and B, then the probability that a category-A response was made was given by $\frac{B_A}{B_A+B_B}$.

Finally, we fit a pair of RB4 models equivalent to the models fit by Schnyer et al. (2009), and are typically used as representatives of the RB class of models by Maddox and colleagues for data from this type of experiment (e.g., Maddox et al., 2004). The rule used in RB Model 4a is that any stimulus shorter than X_1 or longer than X_2 is called A or C, respectively. Any line of intermediate length is either called B or D, depending on whether its orientation is steeper or flatter than Y, respectively. In RB Model 4b, the rules are simply reversed for length and orientation: Any stimulus steeper than Y_2 or flatter than Y_1 is called B or D, and lines in between are either A or C if they are longer or shorter than X. Note that these models are a special case of RB Models 1a and 1b, where either $Y_1 = Y_2$ or $X_1 = X_2$, respectively.

Figure 4 provides schematic illustrations of prototypical examples from each main set of RB models. The color of each region represents the category into which the stimulus would be placed. Regions with two overlapping colors represent regions in which a guess (possibly biased) must be made between the two categories corresponding to those labels. Finally, white regions represent regions in which a participant must guess (sometimes biased) among all four categories.

The RB models from Sets 1 and 2 without response bias require six free parameters (s_X, s_Y, X_1, X_2, Y_1, and Y_2). The models from these sets that included response bias estimated a further three parameters, for a total of 9 (s_X, s_Y, X_1, X_2, Y_1, Y_2, b_{A}, b_{B}, b_{C}). The RB models from Sets 3 and 4 each used five free parameters (s_X, s_Y, and either X_1, X_2, and Y, or Y_1, Y_2, and X).

Information-integration models. We also fit a number of models from two sets of II models: prototype and exemplar. These II models all make the approximate prediction shown by the diagonal lines in Figure 1. However, for completeness, we also fit the diagonal-lines II model to the data.

In formalizing the prototype and exemplar models, we followed the development provided by Nosofsky (1986, 1987). In the prototype models, participants are assumed to make categorization decisions on the basis of the similarity between a presented stimulus and the prototypical member of each category. We assumed that the prototype for each category was the average of all of the items from that category. Note that prototypes were defined by the correct responses and not the participant's responses. The psychological distance, d, between stimulus i and the prototype from category J is given by $d_{iJ} = (\sum_m w_m(|x_{im} - x_{Jm}|)^r)^{1/r}$, where w_m is the attention paid to stimulus dimension m. We fit two versions of the model, one which set r to be 1, corresponding to the city-block distance metric, and another which assumed that r was 2, using the Euclidean distance metric. The distances were converted to similarities by taking $s_{iJ} = e^{-cd_{iJ}}$, where c is a sensitivity parameter. Finally, the probability of responding category J for stimulus i is given by $\Pr(J|i) = \frac{s_{iJ}}{\sum_J s_{iJ}}$.

We also fit versions of the prototype model that assumed participants would be biased toward particular responses. To do this,

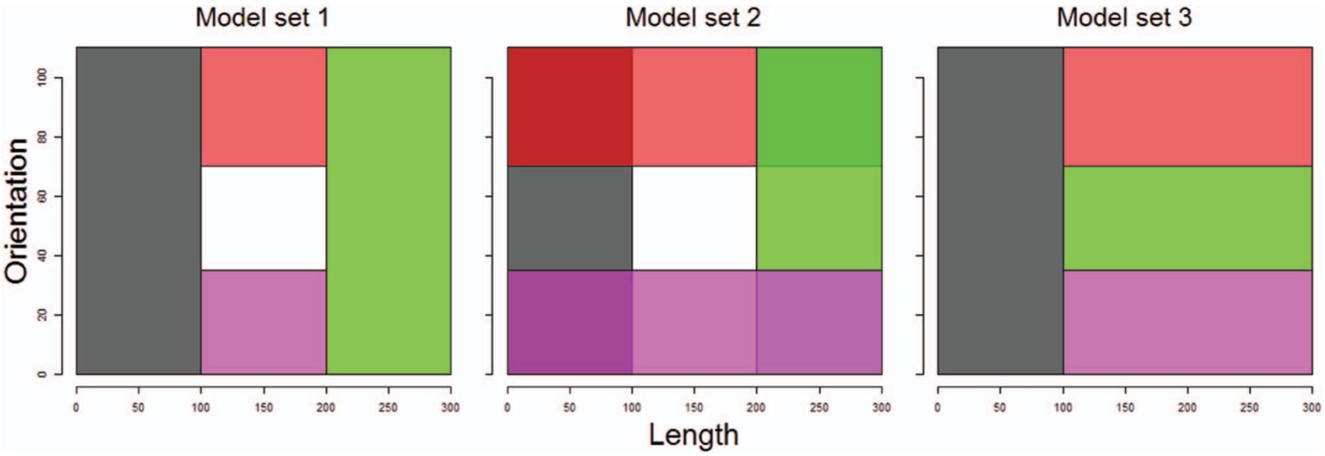

Figure 4. Schematic illustration of three of the four main sets of RB models we fit to data. The color of the region defines the category into which a stimulus from that region would be classified (A = black, B = red, C = green, D = violet). Model Set 1 has three main versions: 1a is shown, 1b is achieved by having the orientation dimension have the simple rule, and 1c is an equal mixture of 1a and 1b. Model Set 2 has only a single main version, depicted in the figure. Both Model Sets 1 and 2 also have subversions with and without biased guessing (see text). Model Set 3 has four versions: 3a is shown, with the others having the other three categories defined by the simple rule. Models 4a and 4b are equivalent to Models 1a and 1b, but with no central guessing area. See the online article for the color version of this figure.

we again assumed three additional free parameters b_{A}, b_{B}, and b_{C}, and set b_{D} to be 1, such that the bias for category J was $B_J = \frac{b_J}{b_A+b_B+b_C+1}$. Now, the probability that category J was the response made for stimulus i was given by $\Pr(J|i) = \frac{B_J s_{iJ}}{\sum_J B_J s_{iJ}}$.

For reasons explained later in our article, we also fitted an extension of the prototype model that included a background-noise parameter (e.g., Stanton & Nosofsky, 2013). In this extended model, the probability of responding category J for stimulus i is given by $\Pr(J|i) = \frac{s_{iJ}+back}{\sum_J (s_{iJ}+back)}$. In brief, the background-noise parameter makes allowance for the possibility that observers will exhibit guessing behavior in regions of the stimulus space that are distant from the prototypes of all categories: When similarity to all prototypes is low, the background-noise constant dominates the response rule, yielding guessing behavior in those regions of the space. To avoid an explosion of different II models, we investigated the potential role of the background-noise parameter only for the case of the unbiased, city-block prototype model.

The exemplar-based models were very similar to the prototype models, except that instead of comparing the similarity between stimulus i and the prototype for category J, the classification is based on the summed similarity between stimulus i and all of the individual members, j, of category J. As such, the probability of responding with category label J is given by $\Pr(J|i) = \frac{S_J}{\sum_J S_J}$, where S_J is the sum of the similarity between stimulus i and all of the individual members of category J. All of the previous formulae for distance and similarity are the same for the exemplar-based model, except that now they are calculated between stimulus i and category member j (i.e., not the prototype). Also, like the prototype models, we fit exemplar-based models that assumed city-block and Euclidean distance metrics, and which did and did not assume response bias. Finally, we also fit a version of the exemplar-based model that assumed an extra γ parameter, which allowed the model to be more or less deterministic, by having the probability of responding be $\Pr(J|i) = \frac{S_J^{\gamma}}{\sum_J S_J^{\gamma}}$ (Nosofsky & Zaki, 2002). Two versions of this γ exemplar-based model, either with or without response bias, were fit to the data.

The prototype and exemplar models without response bias had just two free parameters (w and c), and adding the background-noise parameter yielded a three-parameter model. The models that included response bias had an additional three parameters (b_{A}, b_{B}, and b_{C}). The γ version of the exemplar-based model had one additional γ parameter.

For completeness, we also fitted the "diagonal-lines" II model to our data (which is typically the sole II representative in the past studies that have attempted to identify RB vs. II strategy use). According to that model, the observer partitions the stimulus space using diagonal lines of the form illustrated in the left panel of Figure 1. However, the slope and y-intercept of each line are allowed to be free parameters. As in the RB models, each stimulus is represented by a bivariate normal distribution, and the probability that a stimulus is classified into a category is given by the proportion of its distribution that falls in the category region defined by the diagonal lines. For simplicity in computing the predictions from the model, we assumed that the standard deviations of the stimulus distributions were the same along dimensions x and y ($s_X = s_Y$). However, to provide the model with appropriate flexibility, the stimulus locations along dimension x were scaled by a free parameter. The diagonal-lines II model has six free parameters: the slope and y-intercept of each diagonal line boundary, a perceptual standard deviation parameter, and the dimension-x location-scaling parameter. We should note that the diagonal-lines II model is closely related to the Euclidean prototype models, but grants those models additional flexibility. According to the Euclidean prototype models, a stimulus will tend to be classified into the category that has the closest prototype. Assuming a Euclidean distance metric, these closest-distance relations are depicted by drawing diagonal lines through the stimulus space. That is, all stimuli that are closer to Prototype A will fall to the A side of a dividing diagonal line, whereas all stimuli that are closer to Prototype B will fall to the other side of that diagonal line. However, whereas in the prototype model the locations of those diagonal lines are determined by the coordinates of the prototypes, in the present diagonal-lines II model the locations are given by freely estimated parameters.

Model analysis. In total, there were 14 RB models, 12 II models, and one baseline "biased-guess" model, which simply assumed that participants would give a potentially biased guess among the four categories for all stimuli. This final model estimated three free parameters (b_{A}, b_{B}, b_{C}; and b_{D} was set at 1). The probability of a category J response for any stimulus in this model was simply $\Pr(J) = \frac{b_J}{\sum_J b_J}$.

Each model produced a set of predicted probabilities that a particular category was chosen for each stimulus. These predicted probabilities give the likelihood of each response made to each stimulus. The overall likelihood of the model is then found by computing the product of these individual-response likelihoods. Each model was fit to data by conducting a computer search for the parameter values that maximized this overall likelihood. Best-fitting parameters were found via a SIMPLEX search, with each search started from multiple start-points.

We began by fitting each of the 27 models to all of the data from each individual's final test block, that is, both the training and transfer trials. For each model's fit to each individual's data we calculated a Bayesian Information Criterion (BIC) score using $BIC = k \log N - 21$, where k is the number of free parameters in the model, N is the number of data points fit by the model, and l is the (maximum) log-likelihood of the model given the data. BIC is smaller as the fit of the model improves, but increases as the model becomes more complex. As such, the model with the smallest BIC value is said to give the most parsimonious explanation of the data.

Table 2 contains the BIC values for the best fitting model from each of the RB and II model classes. Inspection of the table reveals that there are many participants for which a member of one of the two classes of models provides a clearly better account of the data than does the other model class. However, there are a reasonable number of cases (Participants 31 to 45) for which the difference between the best-fitting RB and II models is small (i.e., less than 5 BIC points).

Table 2
BIC Values for the Best Fitting RB and II Model

	RB		II				RB		II		
Participant	BIC	Model	BIC	Model	Acc	Participant	BIC	Model	BIC	Model	Acc
1	250	2-b	**191**	**P-e**	0.91	32	254	1b-b	251	P-B-cb	0.84
2	232	2-b	**182**	**P-B-e**	0.93	33	238	2-b	235	P-e	0.89
3	260	4a	**210**	**P-B-e**	0.86	34	277	2-b	275	P-e	0.85
4	250	2-b	**202**	**P-B-e**	0.91	35	191	4a	190	P-e	0.86
5	270	4a	**221**	**E-B-cb**	0.85	36	313	1c-b	312	P-B-cb	0.85
6	262	2-b	**222**	**P-e**	0.91	37	319	2-b	319	P-B-e	0.82
7	245	2-b	**206**	**P-B-e**	0.89	38	424	1a	423	P-e	0.57
8	322	1b	**287**	**P-e**	0.84	39	355	1a-b	354	E-B-cb	0.77
9	243	2-b	**208**	**P-e**	0.87	40	251	2-b	250	P-e	0.87
10	186	2-b	**154**	**P-B-e**	0.89	41	258	2-b	259	P-e	0.85
11	296	2-b	**264**	**E-B-cb**	0.82	42	406	3a	407	P-B-e	0.68
12	450	3a	**420**	**diag**	0.71	43	205	2-b	207	E-B-e-g	0.92
13	270	2-b	**241**	**P-e**	0.86	44	247	2-b	250	P-e	0.87
14	250	4a	**224**	**E-B-cb**	0.81	45	**257**	**3a**	262	P-e	0.77
15	294	1c	**268**	**P-e**	0.81	46	**308**	**1c-b**	313	P-back	0.8
16	268	4a	**246**	**E-B-cb**	0.88	47	**273**	**2-b**	280	diag	0.82
17	464	4a	**447**	**P-cb**	0.48	48	**199**	**4a**	220	P-B-e	0.82
18	278	1c-b	**260**	**P-B-e**	0.86	49	**584**	**3a**	607	diag	0.36
19	417	1b	**399**	**P-B-e**	0.66	50	**330**	**2-b**	351	E-B-cb	0.68
20	273	2-b	**257**	**P-e**	0.88	51	**223**	**3a**	255	E-B-cb-g	0.84
21	332	4a	**316**	**E-B-cb**	0.76	52	**273**	**3a**	309	diag	0.83
22	358	1a	**344**	**P-cb**	0.6	53	**518**	**1a-b**	556	diag	0.55
23	244	2-b	**229**	**diag**	0.87	54	**244**	**2-b**	290	E-B-cb	0.8
24	389	1a	**375**	**diag**	0.53	55	**293**	**3a**	343	diag	0.74
25	286	1a	**275**	**P-cb**	0.62	56	**272**	**2-b**	329	diag	0.76
26	301	2-b	**290**	**E-B-cb**	0.8	57	**452**	**2-b**	513	diag	0.54
27	309	2-b	**300**	**P-B-cb**	0.8	58	**210**	**2-b**	275	E-B-e	0.83
28	457	1a	**452**	**E-B-cb**	0.52	59	**249**	**3a**	347	diag	0.74
29	261	2-b	**256**	**E-B-cb**	0.82	60	**169**	**3a**	298	diag	0.78
30	281	1c-b	**276**	**E-B-e**	0.77	61	**160**	**3a**	287	diag	0.84
31	205	4a	201	P-e	0.85	62	**417**	**1a-b**	498	diag	0.55

Note. Participants are sorted based on how much better they were fit by II models. The best-fitting model is bolded in cases where participants are clearly favored by one model. Average accuracy for training stimuli in the final, 5th block is reported under the heading ‘Acc’. Rule-based models are defined in text. II models are defined using: P = prototype model; E = exemplar model; diag = diagonal bound model; B = bias; e = Euclidean distance metric; cb = city-block distance metric; g = gamma; back = background noise.

Table 2 also reports each participant’s average accuracy for training stimuli in the final block of trials. The average proportion of correct responses for participants identified as using the correct strategy (II) was 0.79 (*SD* = 0.12). Participants identified as using rules performed less well on average ($M = 0.76$, $SD = 0.14$). The difference in accuracy, however, is inconclusive according to the results of a default Bayesian *t* test, with the Bayes factor suggesting that the difference is only 2.5 times more likely under the alternative than the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009). As we speculated earlier, it seems that RB strategies provided a viable means of learning to categorize the stimuli from this II structure.

The size of the difference between two BIC scores reflects the degree to which one model fits better than another model. Rather than simply consider which model had the smallest BIC, we can instead take the size of the difference between BIC scores into account. Wagenmakers and Farrell (2004) outline a method by which BIC scores can be turned into BIC weights. These weights can be interpreted as the probability that a particular model generated the observed data. First, we calculated the difference between the BIC for each model and the best-fitting model, ΔBIC. Then, the relative likelihood for the *i*th model is calculated via $L(M_i|data) = e^{-\frac{1}{2}\Delta BIC}$. Finally, the BIC weights, w_i, are calculated using $w_i = \dfrac{L(M_i|D)}{\sum L(M_i|D)}$.

Figure 5 contains a heat map of all of the model probabilities for each of the 27 models (columns) for each individual participant (rows). White cells correspond to a model probability of 1, red cells correspond to a model probability of 0, and the closer the color of the cell is to white, the closer the probability of the model is to 1. The left 14 columns of the figure correspond to RB models, the next 12 columns correspond to II models, and the final column corresponds to the biased-guess model.

Figure 5 shows that there are considerable individual differences in which model gives the best account of data. Although the majority of participants are best fit by II models, many are best fit by an RB model. There were two particularly successful RB models: the RB Model 2 with bias, and the RB Model 3a. Also interesting is that only one participant was best fit by an RB model from Set 4, that is, those that are usually applied to data from this task (e.g., Schnyer et al., 2009). Of the II models, there is much less certainty as to whether any one or two models give the best account of behavior. Instead, model probabilities are divided

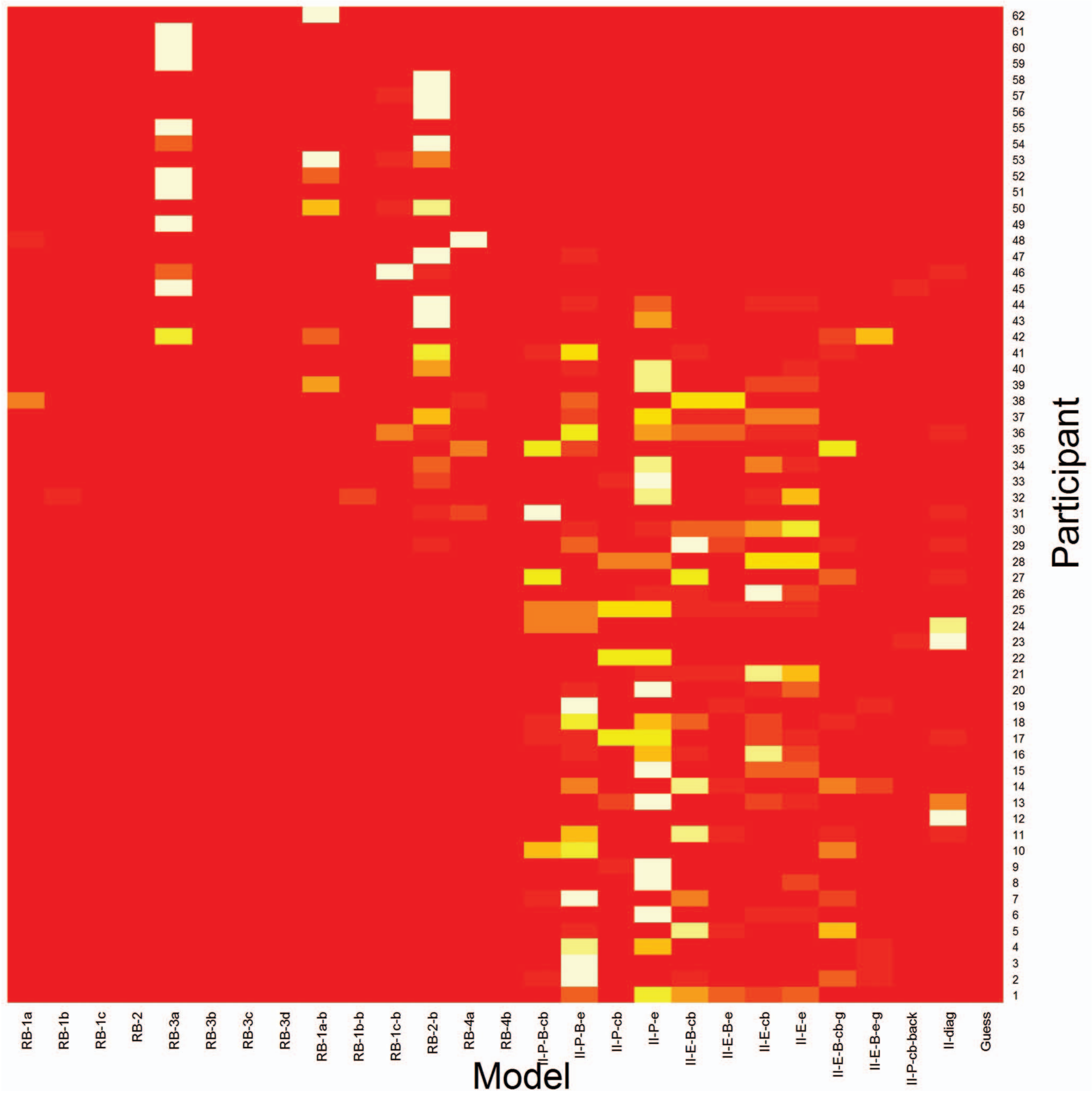

Figure 5. Heat map of BIC model probabilities for each of the models (columns) for each participant (rows). White implies a model probability of 1 and red a model probability of 0. The models are defined using the same notation as in Table 2. Participants were sorted as in Table 2. See the online article for the color version of this figure.

among many of the II models. It is also worth noting that no participant in our study was identified as guessing randomly, which is contrary to what is usually observed (e.g., Maddox et al., 2010; Schnyer et al., 2009).

To seek evidence that participants best fit by an RB model did indeed appear to be using rules, we plotted the classifications made by participants against the predictions of the best-fitting RB models. We chose eight participants who were exemplary of how well the RB models did indeed fit some participants (see Figure 6). The top row of participants in the figure are those that were best fit by an RB model from Sets 1 or 2, and the bottom row are those that were best fit by an RB model from Set 3.

To take one example, consider the responses produced by the participant in the top row, leftmost panel of Figure 6. It is straight-

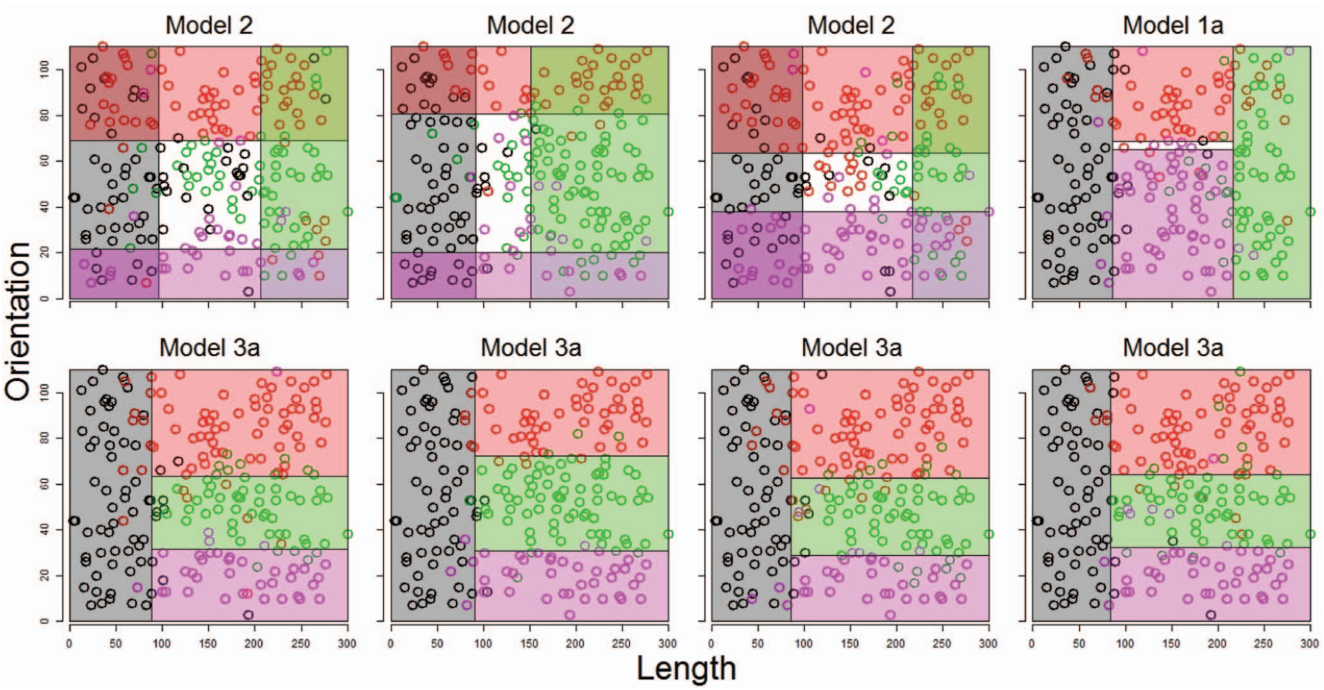

Figure 6. The best fitting RB models for eight exemplary rule-using participants. The category chosen by the participant for each stimulus is shown by the color of the circle. The corresponding rule boundaries of the best-fitting RB model is shown by the shaded area. The color of the shaded area corresponds to the category chosen. When two shaded areas overlap, then the participant had to guess between those two categories. See the online article for the color version of this figure.

forward to see why RB Model 2 fit these data so much better than did the competing II models: Namely, the classifications for stimuli in the corners of the stimulus space tend to be unrelated to the location of these stimuli within these regions, but are instead distributed randomly. The RB model can handle this pattern, whereas the II models must predict an orderly relation between location and categorization: stimuli closer to the prototypical item from a category (or closer to all other stimuli in that category) are more likely to receive that particular label. The same pattern holds for the other participants who are best fit by RB Model 2. Note that Figure 6 is not intended to reflect the "average" fit of the model, but simply shows that some participants were clearly employing a RB strategy.

An II-theorist might argue that once a stimulus is sufficiently far from the prototypes of all categories, similarity to the prototypes no longer governs responding, and the observer guesses randomly. The prototype model with background noise was intended to formalize that possibility. However, the background-noise model almost always yielded worse BIC scores than did even the standard prototype model. The reason is that although responses tend to be distributed randomly within the corners of the stimulus space for these subjects, they are always restricted to the two categories indicated by the two competing rules. By contrast, the background-noise model predicts that when a stimulus is far from all categories, guessing will be random among all the categories.

Finally, inspection of the results from the participant in the top row, rightmost panel of Figure 6, as well as all participants in the bottom row, reveals that RB Models 1a and 3a neatly partition these classification response profiles. It is easy to see why the RB models provided a dramatically better fit to these subjects' data than did any of the II models.

Thus far we have focused on which particular model has best fit an individual, but one may also be interested in the evidence for the entire class of either RB or II models. We outline two potential approaches to answering this question, but note that both have their advantages and disadvantages. One option is to select the best fitting model from each of the RB and II classes of models and calculate BIC weights for those two model variants, as we illustrated earlier. A potential disadvantage to this approach is that a class of models with more variants will have a higher probability of being selected before we observe any data. An alternative approach could punish classes of models with more variants, by calculating BIC weights such that the classes have an equal probability before any data are observed. That is, we define the BIC weight, $w_i = \dfrac{k_i L(M_i|D)}{\sum k_i L(M_i|D)}$, where k_i is the number of models in the model class to which it does not belong divided by the total number of models. The BIC weights for each model in a given class are summed to create an overall probability of each class. A downside to this approach is that one could propose a large number of unlikely models of a particular class so as to underweight the evidence for any particular variant from within that class.

As it turned out, for our data, the two approaches yielded the same conclusions. We report the results based on the second approach outlined above, but note that all of what follows is consistent with the approach where class comparison is based only on the best-fitting model of each class (the information necessary to carry out this comparison is reported in Table 2).

The resultant model-class probabilities are plotted for each participant in the top left panel of Figure 7. The probabilities show that 40 of the 62 participants were better fit by the II class of models. One participant was fit better by neither model, leaving 21 participants whom were fit better by an RB model. However, it is clear from Figure 7 that the evidence for RB and II models was not always definitive. For example, the probability that the best-fitting model class was either RB or II was below 0.9 for 14 participants, or almost a quarter of our sample. Considering only those participants for whom model probabilities were greater than 0.9 leaves 18 RB and 30 II participants.

We find a relatively large number of RB participants, despite using an II category structure. Two aspects of what we have done

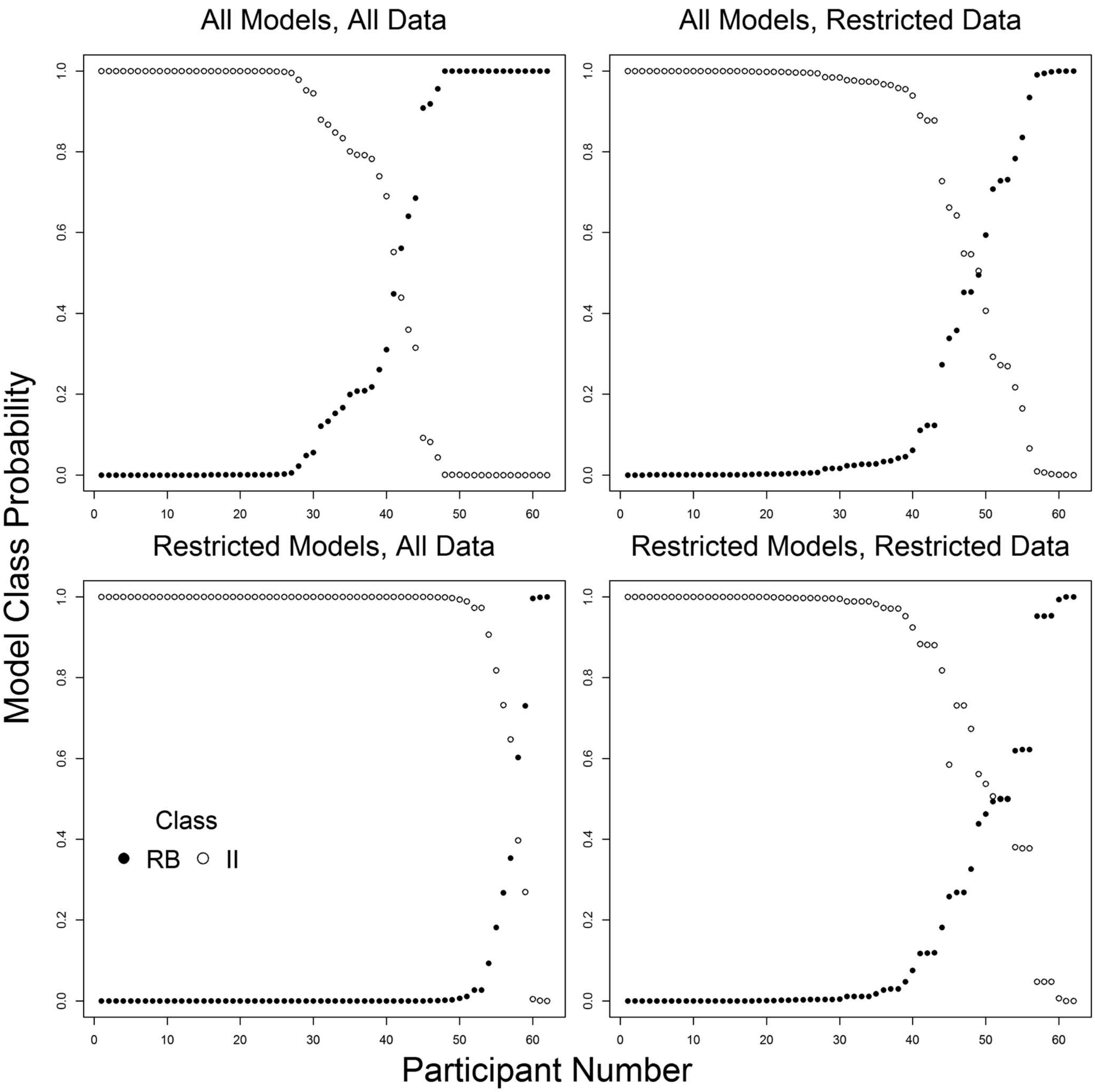

Figure 7. Probabilities for the class of RB (black dots) and II (white dots) models for each individual. The participant numbers are the same as those in Figure 5. The probabilities for each participant sum to 1, because the guessing model was assigned effectively zero probability for all participants.

thus far differ from standard approaches: First, we included additional diagnostic transfer stimuli, and second, we allowed for the adoption of a much larger range of rule-based strategies for classifying stimuli. In the next section, we explore the impact of each of these changes to standard practice.

Alternative Analyses

In previous analyses of the category structure we used, the standard approach has been to compare an II model to the two RB models from Set 4. While these models do represent a plausible set of rules that might be used to classify stimuli from this experiment, many other possible rules also seem plausible. Indeed, based on our analyses, it seems clear that many of our participants did indeed employ some of these alternative rules (see Figure 6). It seems possible that restricting the range of RB models fit to data would also reduce the proportion of participants who were correctly identified as using an RB strategy.

It is also possible that the category structure itself does not provide the ideal amount of information with which to identify participants as using an RB strategy. It seems possible that removing our transfer stimuli and considering only the stimuli that have been used in previous applications might result in more participants being classified as using an II strategy.

To test these ideas, we repeated the above analysis but with some modifications. First, we repeated our model-based analysis using only the RB Models 4a and 4b, the diagonal II model, and the random-guess model (i.e., a *restricted models* analysis). Second, instead of including both training and transfer stimuli, we fit the models to the data from the training stimuli in the test block only (i.e., we excluded the transfer stimuli from analysis—a *restricted data* analysis). There are, therefore, four possible analyses we might conduct: restricted models and restricted data, which would be the standard method for analyzing data from this type of experiment; restricted models but full data; restricted data but full set of models; and the full set of models and full data. The results of this final combination have already been reported, so we now look at the results of the other three analyses.

We computed the model class probabilities for each of the three remaining analyses, and the resultant model-class probabilities are plotted in Figure 7. When both the models and data were restricted, only nine of the 62 participants were better fit by the RB class of models. This number is clearly much smaller than the 21 participants in the full analysis. The number of RB strategy users decreased to just five participants when the full data set was used with the restricted set of models. This result highlights the potential danger of considering only a couple of restricted RB models in the analysis.

Even when the restricted data set is used, it is still useful to include the full set of RB models. When the full set of models is used, the number of RB participants is 13, more than the nine RB participants when the smaller set of models is used. However, that this final number falls short of the 21 RB participants in the full analysis is evidence that the combination of a rich data set and a broader set of RB models has additional benefit when identifying participant classification strategies. Finally, note that there are still many participants for whom the probability of the best-fitting model class falls well short of providing strong evidence (i.e., model class probability of less than 0.9).

Discussion

The take-home message from our work is straightforward—the conclusions about classification-strategy use in experiments that test the Figure 1 II category structure depend dramatically on the analysis. Although our work considered only this single category structure, it is an important one because the Figure 1 four-category stimulus space is widely considered a benchmark II structure. Furthermore, it has been used extensively in numerous studies that investigate the extent to which different participant and patient groups use RB versus II classification strategies. Beyond this particular structure, we believe that the implications of our conclusions and subsequent recommendations can be generalized to numerous pursuits to identify the classification strategies of participants.

Our analysis led to the conclusion that 30 participants used an II strategy (48%), more than the 18 using an RB strategy (29%), and that 14 participants were not clearly preferred by either model class (23%). This result is reassuring considering that the task we used was an II task. However, our evidence pales in comparison to the support for an II strategy when we use the *standard* approach (see Maddox et al., 2004, 2010; Schnyer et al., 2009), where 55 participants are identified as II participants (89%), leaving only seven RB participants (11%). Our approach differs from the standard one in three important ways—we used a more comprehensive category space, considered more rules, and used a better technique for doing model selection. The following discussion will focus on why each of these factors is important.

Category Space

In the standard approach, only training stimuli are included in the design. However, for the Figure 1 category structure, this approach yields very few stimuli that provide sharp contrasts between the predictions of plausible RB and II models. In our design, we included nonreinforced transfer stimuli in diagnostic regions of the stimulus space, thereby allowing clearer discrimination between the alternative models. Inspection of the response profiles in these diagnostic regions provided strong evidence that, despite the use of an II category structure, many of our participants did indeed use RB strategies rather than II ones.

We conjecture that many participants used rules in this task only because the category space made it a viable strategy. If participants had received feedback for stimuli in both the training and transfer regions for the whole experiment (with correct answers defined in terms of an information-integration model), such rule-based strategies would not have yielded near perfect performance. Such corrective feedback may drive participants to adopt a classification strategy closer to that of information integration models. Therefore, our recommendation is that if one wishes to provide convincing evidence that II strategies are being used, II category spaces should be designed such that simple rule-based strategies cannot yield high performance levels.

Rule-Based Models

Standard practice for this paradigm is to consider two possible rule-based models for classifying stimuli. We considered 12 other models implementing alternative rules that we thought were a priori plausible for this category space. Approximately 30% of our participants were identified as being highly likely to have utilized one such rule. When we fit only the standard pair of RB models, only 6% to 11% of participants (depending on whether the full or restricted data set was used) were identified as using an RB strategy. Such a dramatic shift in conclusions makes clear the need to consider a range of possible rule-based strategies.

It is interesting that not one participant was identified as using a random-guess strategy. Generally, using the standard approach, a subset of participants is identified as using a random-guess strategy (e.g., Maddox et al., 2009). In at least some of these cases, we suspect that these participants were employing an ineffective rule-based strategy, and were incorrectly identified as guessing randomly because that particular RB model was not considered.

Model Selection

Standard practice is to simply count up the number of participants who were better fit (in terms of a fit statistic such as BIC) by each model or model class. Our results highlight why this practice is inappropriate. The issue is that not all differences between models are equivalent: a BIC difference between two models of 1 provides less evidence than a difference of 10. We recommend transforming raw BIC scores into BIC weights and model probabilities, as this quantifies the amount of evidence for models (Wagenmakers & Farrell, 2004). In our data, almost a quarter of our participants were not clearly identified as using either an II or RB strategy.

In the analyses presented here, we assigned individuals whose model probabilities were greater than 0.9 the status of "clear evidence" for a particular model or model class. A criterion of 0.9 is equivalent to a Bayes factor of 9 (i.e. $\frac{0.9}{(1-0.9)}$, indicating that the data were 9 times as likely under one model than the other. It is important to note that the aim of this criterion was simply to provide a reminder that not all evidence is clear, and to highlight the potential issues with ignoring the degree of support for models. In practice, the use of such binary criteria reduces the information we obtain from our data. If participants must be classified as RB or II users, then we would suggest that this only be done for participants for whom the evidence for a particular model is strong, and that the continuous model probabilities be presented alongside any such classification, so as to avoid reducing the information present in the data.

Implications for Identifying Categorization Strategy

The implication of our results for the security of conclusions about categorization capabilities of different patient or participant groups is clear. The inability to draw inferences on the basis of raw accuracy data places model-based analysis at the forefront of theorizing. Thus executing the model-based analyses as carefully as possible is crucial.

As an example, the Schnyer et al. (2009) paper discussed in the introduction drew far-reaching conclusions about the role of the ventromedial prefrontal cortex in learning II and RB tasks. These conclusions were drawn on the basis of 13 patients and 11 controls who learned the Figure 1 II structure. Schnyer et al. (2009) argued for poor strategy selection in patients with lesions, only six of 13 of whom were classified as II users, by contrasting them with the 11 control participants all classified as II users. Our results suggest, however, that many of the control participants may have been using rules as well. Thus, the better performance of the controls may suggest that they simply used more effective rules, rather than that a specialized II neural system was spared. More generally, given our demonstrations of the difficulties in telling apart RB from II strategies using standard practice, the sample sizes seem too small to allow any strong conclusions.

Implications for Multiple Categorization Systems

Although the focus of the present work concerned strategy identification in RB and II tasks, the results are also relevant to the issue of "single systems" versus "multiple systems" in categorization. A central theme of COVIS theorists has been to argue in favor of the existence of multiple systems of categorization—specifically, an explicit system that constructs verbalizable rules and an implicit system based on procedural learning. By contrast, theorists such as Nosofsky (1986; Nosofsky & Johansen, 2000) have suggested that a single exemplar-similarity system that makes allowance for selective-attention processes may be sufficient. The present results clearly challenge the exemplar-based single-system view: For many of our participants, RB models provided clearly better accounts of the patterns of classification responding than did II models such as exemplar models. Another example of recent evidence that points strongly to RB forms of classification are patterns of response-time (RT) data that are well captured by RB models but not by exemplar models (e.g., Fific, Little, & Nosofsky, 2010; Lafond, Lacouture, & Cohen, 2009; Little, Nosofsky, & Denton, 2011).

However, the debate between multiple-system theorists and single-system theorists is more nuanced than simply whether multiple systems of categorization exist. A key issue in the debate concerns what constitutes clear-cut diagnostic evidence that can be used to discriminate between these contrasting perspectives. For example, a major vehicle that COVIS theorists have used to support the multiple-systems thesis involves the demonstration of a variety of dissociations in which the manipulation of certain experimental variables has differential effects on performance involving II versus RB tasks. By contrast, researchers who have questioned the COVIS theory have raised concerns about various confounds in these experimental designs and whether the dissociations are indeed diagnostic of multiple systems (for reviews, see Dunn, Kalish, & Newell, 2014; Newell, in press; Newell et al., 2011). Here, we present clear evidence that participants learn to categorize in more than one particular way, and do not have to rely on dissociations to do so.

On the other hand, it is not necessarily the case that the existence of multiple strategies for solving categories necessitates the existence of two distinct systems for category learning. Further, our results do not speak to any of the specific assumptions made in COVIS about those two systems. For example, rule-base strategies in the COVIS model are assumed to be verbalizable and implemented by participants in an explicit fashion. We provide no evidence that all of the rule models we fit are readily verbalizable. Thus while the present evidence suggests participants adopt a mix of II and RB strategies when learning an II task, it is agnostic with regard to the hypothesized markedly different properties of the system(s) that COVIS theorists propose underlies such learning.

Limitations and Model Recovery Simulation

One of the major advances in the present work was our use of transfer stimuli to help diagnose the classification strategy being used by the individual observers. A potential problem involving the use of transfer stimuli, however, involves the logical possibility that their introduction may cause an observer to switch the classification strategy they would otherwise have used. One reason why we continued to present a high proportion of training stimuli with feedback during the transfer phase was to try to maintain consistent response strategies across blocks. Still, we acknowledge the possibility that strategies may have changed. To partially address this possibility, we analyzed results for the training stimuli presented during Blocks 4 and 5. Averaged across participants, the proportion of cases in which the response given to a training stimulus during Block 5 was the same as the response given to that training stimulus during Block 4 was 0.77. Of course, even if the classification strategy remained exactly the same, some proportion of responses would switch across blocks due to noise in the stimulus representations or the response rule—indeed, the same proportion calculation was 0.79 when comparing Blocks 3 and 4. This concordance in the proportion of responses that stayed the same is consistent with the hypothesis that any changes in classification strategies across the training and transfer blocks were not dramatic. Nevertheless, future work should address the strategy-change hypothesis in greater detail. In the meantime, we maintain our view that our present findings point to the dangers of assessing classification strategies using standard practice.[3]

We used the commonly employed BIC to do our model selection. There are of course a number of alternative methods. The closely related Akaike Information Criterion (AIC), for example, is also often used to decide between models of category learning. AIC applies a less severe penalty for extra parameters than BIC, and as such prefers more complex models. When we repeat our analyses using AIC instead of BIC, we find that even more participants are better accounted for by RB strategies. In particular, we find that the number of people clearly using an RB strategy increases from 18 to 34. The number of II users reduces from 30 to 18, leaving 10 people not clearly preferred by either class of models. We decided to focus on the BIC results in our main report in order to be conservative with respect to our conclusions about the use of RB strategies for the Figure 1 II task.

Both AIC and BIC punish model complexity by counting the number of free parameters, and can therefore be criticized for failing to take into account the functional-form complexity of models (Myung, 2000). In our view, the method we used here is better than standard practice, as differences in BIC are transformed into model probabilities, giving more graded evidence for models. However, future work should look to employ model selection methods that appropriately take into account model complexity such as Deviance Information Criterion, Bayes Factors, or Minimum Description Length (Kalish, Newell, & Dunn, 2014; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Regardless of whether we use AIC or BIC, neither approach takes into account that the class of RB models was likely more flexible than the class of II models. That is, the various RB models we fit seem more different from one another than the II models, so the overall number of data patterns that the RB class can produce is greater than that of the II models.

We conducted a small model-recovery simulation to investigate whether this unbalanced flexibility is likely to have led to a bias against selecting the less flexible II models. We simulated 62 data sets from the Euclidean prototype model with bias. The best-fitting parameters from each individual were used to simulate the data sets, and each simulated data set was of the same sample size and used the same stimuli as the original data.

We then subjected these simulated data sets to the same analysis as was performed on the real data, fitting all 27 models and calculating model probabilities. The model class probabilities for these 62 simulated data sets was such that not one individual was better fit by the RB model class. In fact, 60 of the 62 participants have a model probability for the RB models of less than 0.1 (i.e., “clear” evidence for the II model). This simulation suggests that participants using an II strategy, such as that modeled by the prototype model, were unlikely to have been misidentified as using an RB strategy.

Despite the results of this model-recovery simulation, we acknowledge that a great deal more work is needed to develop principled techniques for comparing classes of models. Whereas measures such as BIC and AIC are often reasonable approximations for comparing individual models, the complexities of model comparison grow when the models are embedded in classes, such as the RB and II classes that we investigated in the present work. As noted earlier in our article, we considered two very different approaches, one based on comparing the best models of each class, and a second based on comparing the average likelihood of the models in each class. For our present comparisons, the two approaches yielded identical results. Furthermore, our inspection of the individual response profiles of the subjects who were well fit by RB models seemed strongly suggestive of RB behavior. Even so, more sophisticated techniques will likely be needed to make progress in comparing the classes of RB and II models.

Conclusion

Comparisons between RB and II models of categorization have played a central role in identifying the classification strategies used across a wide variety of experimental conditions and by varieties of patient groups with neurological disorders. We have provided evidence that many of those comparisons may have been compromised by limitations in a) the diagnosticity of the paradigms used for comparing the models, b) the range of models considered, and c) the techniques of model-selection that were used. In drawing

[3] Note that a model-based analysis of the difference between Blocks 4 and 5 would be of limited utility. As we have shown, analysis of restricted data (only the training stimuli) leaves little precision for identifying categorization strategy.

inferences about the nature of neurological patient disorders, the stakes are even higher than in other forms of basic psychological research. Future work is this highly significant area should probably focus on the testing of category structures with a smaller number of plausible model candidates and designs that yield sharp qualitative distinctions between the predictions from the general model classes.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481. http://dx.doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*, 204–210. http://dx.doi.org/10.1016/S1364-6613(00)01624-7

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

Ashby, F. G., & Spiering, B. J. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews, 3*, 101–113. http://dx.doi.org/10.1177/1534582304270782

Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby (2014). *Psychonomic Bulletin & Review, 21,* 947–954. http://dx.doi.org/10.3758/s13423-014-0637-y

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. http://dx.doi.org/10.1037/a0027867

Ell, S. W., Marchant, N. L., & Ivry, R. B. (2006). Focal putamen lesions impair learning in rule-based, but not information-integration categorization tasks. *Neuropsychologia, 44,* 1737–1751. http://dx.doi.org/10.1016/j.neuropsychologia.2006.03.018

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117,* 309–348. http://dx.doi.org/10.1037/a0018526

Huang-Pollock, C. L., Maddox, W. T., & Tam, H. (2014). Rule-based and information-integration perceptual category learning in children with attention-deficit/hyperactivity disorder. *Neuropsychology, 28,* 594–604. http://dx.doi.org/10.1037/neu0000075

Kalish, M., Newell, B. R., & Dunn, J. C. (2014). *When more is better: Higher working memory capacity facilitates perceptual category learning*. Manuscript submitted for publication.

Lafond, D., Lacouture, Y., & Cohen, A. L. (2009). Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review, 116,* 833–855.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1–27. http://dx.doi.org/10.1037/a0021330

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

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

Maddox, W. T., Glass, B. D., Zeithamova, D., Savarie, Z. R., Bowen, C., Matthews, M. D., & Schnyer, D. M. (2011). The effects of sleep deprivation on dissociable prototype learning systems. *Sleep, 34,* 253–260.

Maddox, W. T., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D. M. (2010). Rule-based and information-integration category learning in normal aging. *Neuropsychologia, 48,* 2998–3008. http://dx.doi.org/10.1016/j.neuropsychologia.2010.06.008

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 294–317. http://dx.doi.org/10.1037/0096-1523.22.2.294

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. http://dx.doi.org/10.1006/jmps.1999.1283

Newell, B. R. (in press). Wait! Just let me NOT think about that for a minute: What role for implicit processes in higher level cognition? *Current Directions in Psychological Science*.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition, 38,* 563–581. http://dx.doi.org/10.3758/MC.38.5.563

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 54, pp. 167–215). Burlington, MA: Academic Press.

Nomura, E. M., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience and Biobehavioral Reviews, 32,* 279–291. http://dx.doi.org/10.1016/j.neubiorev.2007.07.006

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. http://dx.doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108. http://dx.doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of “multiple-system” phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375–402.

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. http://dx.doi.org/10.3758/BF03193227

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940. http://dx.doi.org/10.1037/0278-7393.28.5.924

Pothos, E. M., & Wills, A. J. (Eds.). (2011). *Formal approaches in categorization*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511921322

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia, 47,* 2995–3006. http://dx.doi.org/10.1016/j.neuropsychologia.2009.07.011

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284. http://dx.doi.org/10.1080/03640210802414826

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1174–1191. http://dx.doi.org/10.1037/a0031670

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. http://dx.doi.org/10.3758/BF03206482

Zaki, S. R., & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: Evidence for multiple systems or task complexity? *Memory & Cognition, 42,* 508–524. http://dx.doi.org/10.3758/s13421-013-0375-9

Received June 5, 2014
Revision received September 14, 2014
Accepted September 22, 2014 ■