Memory & Cognition
2010, 38 (7), 916-927
doi:10.3758/MC.38.7.916

Classification response times in probabilistic rule-based category structures: Contrasting exemplar-retrieval and decision-boundary models

ROBERT M. NOSOFSKY AND DANIEL R. LITTLE
Indiana University, Bloomington, Indiana

Experiments were conducted to contrast the predictions from exemplar models and rule-based decision-boundary models of perceptual classification. Observers classified multidimensional stimuli into categories that could be described in terms of easily verbalized logical rules. The critical manipulation was that some pairs of stimuli received probabilistic feedback, whereas other control pairs received deterministic feedback. Despite the probabilistic feedback, the probabilistic pairs and the deterministic pairs were the same distance from ideal-observer, rule-based decision boundaries. Across two experiments with varying category structures, observers classified the probabilistic pairs with slower response times (RTs) and lower accuracies than the comparison deterministic pairs. The effects were relatively long term, extending into test blocks in which all feedback was withheld. The results were as predicted by exemplar models, but challenged models that posit that RT is a function solely of the distance of a stimulus from rule-based boundaries. The studies add considerable generality to previous ones and suggest that, even in domains involving rule-based category structures, exemplar-retrieval processes play a significant role. Supplemental materials related to this article may be downloaded from http://mc.psychonomic-journals.org/content/supplemental.

Two fundamental issues in cognition are the manner in which people represent categories of perceptual objects in memory and the decision processes that they use to make classification judgments. One modern approach to addressing these issues is to develop and test formal models of perceptual classification that account not only for classification choice probabilities, but also for response times (RTs). By testing models on their ability to account for the time course of classification, deeper insights may be achieved into the nature of perceptual category representation.

Two of the major models for accounting for multidimensional classification RTs and choice probabilities are exemplar models and decision-boundary models (e.g., Ashby & Maddox, 1994; Cohen & Nosofsky, 2003; Lamberts, 1995, 1998, 2000; Maddox & Ashby, 1996; Nosofsky & Palmeri, 1997). According to exemplar models, people represent categories by storing individual exemplars in memory, and they classify objects on the basis of their similarity to the stored exemplars. By contrast, according to decision-boundary models, people form decision boundaries to divide a multidimensional stimulus space into response regions. Anytime a stimulus is perceived to lie in Region A, the observer emits a Category A response. Hybrid models have also been proposed that assume multiple forms of category representation (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994; Vandierendonck, 1995). However, the present work is motivated primarily toward the testing of models that assume single-representation systems.

Although exemplar and decision-boundary models are vastly different in underlying spirit, they make remarkably similar predictions of classification RTs and choice probabilities. According to decision-boundary theory, observers should be faster and more accurate at classifying stimuli that lie far from the decision boundary than those that lie close. A fundamental axiom of decision-boundary theory is that perception is a noisy process (Ashby & Townsend, 1986). Because of perceptual noise, stimuli that lie close to the decision boundary have a greater chance of being misperceived as falling to the incorrect side of the boundary than do stimuli that lie far. Thus, accuracy should be greater for stimuli that lie far from the boundary. Likewise, the farther away that a stimulus is from the decision boundary, the easier it should be to evaluate its location relative to the boundary. Thus, stimuli that lie far from the boundary are classified more rapidly than are stimuli that lie close. Indeed, so fundamental is this presumed relation that it has been coined the *RT–distance hypothesis* (e.g., Ashby, Boynton, & Lee, 1994).

A representative from the alternative class of exemplar models is the exemplar-based random-walk (EBRW) model (Nosofsky & Palmeri, 1997). According to the EBRW model, when a test item is presented, all category exemplars stored in memory race to be retrieved (Logan,

R. M. Nosofsky, nosofsky@indiana.edu

1988). The rates at which the exemplars race are determined jointly by their strength in memory and by their similarity to the test item. The exemplar that wins the race on any given step is retrieved and enters into a random-walk decision process. In a two-category situation, the process is formalized as follows: First, there is a random-walk counter with an initial value of 0. The observer establishes criteria representing the amount of evidence needed to make either a Category A response (+A) or a Category B response (−B). Suppose that exemplar *x* wins the race on a given step. If *x* belongs to Category A, the counter is increased by unit value in the direction of +A, whereas if *x* belongs to Category B, the counter is decreased by unit value in the direction of −B. If the counter reaches either criterion +A or −B, the appropriate response is made. Otherwise, a new race is initiated, another exemplar is retrieved, and the process continues. Classification decision time is determined by the time required to complete the random walk.

A major conceptual prediction from the EBRW model is that the most rapid and accurate classification decisions should be made for items that are highly similar to the exemplars of their own category and dissimilar to the exemplars of the alternative category. Under such conditions, each retrieved exemplar tends to come from the same category, so the random walk marches consistently toward a single criterion.

An important corollary of this conceptual prediction is that the EBRW model tends to predict the RT-distance effect (despite the fact that the model does not posit the explicit formation of a decision boundary). In general, in most classification designs, stimuli that lie far from the presumed decision boundary are highly similar only to exemplars of their own category and are dissimilar to exemplars of the contrast category. Thus, the random walk marches consistently to the correct criterion, leading to highly accurate responding and short RTs. By contrast, stimuli that lie close to the boundary are similar both to exemplars of their own category and to exemplars of the contrasting category. Thus, exemplars from both categories tend to be retrieved, the random walk meanders back and forth, and responding is error prone and slow.

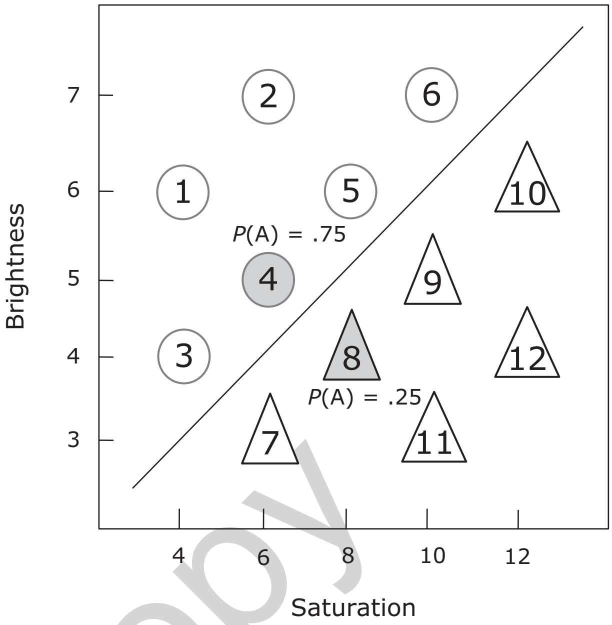

Figure 1. Schematic illustration of the category structure and design tested by Nosofsky and Stanton (2005). Stimuli enclosed by circles represent members of Category A. Stimuli enclosed by triangles represent members of Category B. In one condition, illustrated in the figure, Stimuli 4 and 8 received probabilistic feedback, with Stimulus 4 receiving Category A feedback on 75% of the trials and Stimulus 8 receiving Category B feedback on 75% of the trials. (In a second condition, Pair 5–9 received probabilistic feedback instead.) All remaining stimuli received deterministic feedback. The diagonal line in the figure is the ideal-observer decision boundary for separating the stimulus space into category response regions.

Developing Contrasts Between Exemplar-Retrieval and Decision-Boundary Models

Nosofsky and Stanton (2005) sought to contrast the EBRW model and decision-boundary accounts of classification RTs. (For related approaches, see Nosofsky & Palmeri, 1997, Experiment 2; Rouder & Ratcliff, 2004; Verguts, Storms, & Tuerlinckx, 2003.) The key idea was to decouple the effects of the distance from the boundary and the similarity to the exemplars on RT. The experimental design is illustrated in Figure 1. The stimuli were a set of Munsell colors of a constant red hue varying in their brightness and saturation. Stimuli enclosed by circles belonged to Category A, whereas stimuli enclosed by triangles belonged to Category B. On each trial, a color would be presented, the participant classified it into Category A or B, and feedback was then provided.

The key manipulation was that feedback assignments for some of the individual stimuli were varied across conditions. In particular, certain stimulus pairs received probabilistic feedback, whereas other control pairs received deterministic feedback. In the example in Figure 1, Stimulus Pair 4–8 was a probabilistic pair: Stimulus 4 received Category A feedback on 75% of the trials but received Category B feedback on 25% of the trials. Likewise, Stimulus 8 received Category B feedback on 75% of the trials, but Category A feedback on 25% of the trials. In the Figure 1 example, Pair 5–9 was a deterministic pair that served as a source of comparison with Pair 4–8. Stimulus 5 received Category A feedback on 100% of the trials, and Stimulus 9 received Category B feedback on 100% of the trials.

According to decision-boundary theory, observers will establish a boundary for partitioning the perceptual space into response regions. In the present example, the ideal-observer boundary is the diagonal line illustrated in the figure. It is critical to realize that the very best an observer can do is to simply classify a stimulus into Category A anytime it falls to the upper left of the boundary and to classify a stimulus into Category B any time it falls to the lower right. There is no way to adjust the boundary to classify more accurately the probabilistic pairs, because the trial-by-trial probabilistic feedback assignments are randomly determined. Therefore, because the members of the pairs are the same distance from the decision boundary, these models predict identical classification RTs (and choice probabilities) for the probabilistic and deterministic pairs.

By contrast, the EBRW model predicts longer RTs and less accurate responding for the probabilistic pairs than the deterministic pairs. In the Figure 1 example, in cases in which Stimulus 4 is presented and tokens of Exemplar 4 are retrieved from memory, 75% of the steps in the random walk will move in the direction of Criterion A, but 25% of the steps will move in the direction of Criterion B. By comparison, presentations of the deterministic pairs will result in more consistent steps of the random walk, leading to shorter RTs and more accurate responding.

Across two experiments, the qualitative predictions from the EBRW model were supported over those from the decision-boundary model. In our view, this support for the EBRW model was intriguing: To reiterate, in the Figure 1 paradigm, an ideal observer would simply ignore the probabilistic feedback and emit a Category A response whenever a stimulus falls to the upper left of the diagonal boundary. However, even after 5 days of testing, RTs for the probabilistic pairs were slower than for the deterministic pairs, suggesting a stubborn influence of exemplar-based retrieval.

Probabilistic Feedback in Rule-Based Category Structures

The primary goal of the present research was to test for the generality of these effects using fundamentally different category structures and types of stimuli. In the perceptual classification and cognitive neuroscience literatures, a major distinction has been drawn between rule-based category structures and information-integration structures (Ashby & Maddox, 2005). Intuitively, rule-based structures are those in which it is easy for an observer to verbalize the optimal strategy for classification. In rule-based problems, observers can perform at essentially optimal levels by establishing separate criteria along each of the multiple dimensions that compose a set of stimuli. Separate, independent decisions are then made about a stimulus's value along each of these dimensions. The separate decisions are then combined to determine whether the verbalizable rule has been satisfied. For example, suppose that the stimuli vary in their size and brightness and that membership in a target category is defined by a conjunctive rule. Then a participant may classify an object into the target category if it is both sufficiently large and sufficiently bright. In this case, the observer establishes a fixed criterion along the size dimension and a fixed criterion along the brightness dimension and classifies an object into the target category only if it satisfies both criteria. A key point (see Figures 2 and 4 for illustrations) is that, for rule-based problems, the decision boundaries that divide the space into response regions are straight lines that are orthogonal to the coordinate axes (Ashby & Gott, 1988).

By contrast, in information-integration category problems, the optimal strategy requires that perceptual information from at least two dimensions be combined prior to any classification decisions. Instead of making separate, independent decisions along each dimension, the decision boundaries for information-integration problems will not be orthogonal to the coordinate axes and will be difficult or impossible to verbalize. An example of an information-integration structure is the Figure 1 structure used in the Nosofsky and Stanton (2005) experiments. Here, the optimal decision boundary is the diagonal line illustrated in the figure.

In our view, a compelling idea is that whereas it may be difficult to form decision boundaries for information-integration structures, it may be far more natural to form decision boundaries for rule-based ones. Indeed, the classic idea that much of classification proceeds by establishing and evaluating logical rules continues to be a central one in cognitive psychology (Ashby et al., 1998; Ashby & Maddox, 2005; Erickson & Kruschke, 1998; Feldman, 2000; Fific, Little, & Nosofsky, 2010; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Lafond, Lacouture, & Cohen, 2009; Nosofsky et al., 1994). Furthermore, researchers have argued that performance on rule-based and information-integration category structures may be mediated by separate cognitive systems and by distinct anatomical structures in the brain. They have also amassed much evidence to support this view (Ashby & Maddox, 2005). For example, a wide variety of behavioral and neuropsychological dissociations have been reported in which different effects of experimental variables are observed for rule-based and information-integration structures.

Accordingly, in the present experiments, our primary goal was to test for effects of probabilistic feedback assignments in situations involving rule-based category structures rather than information-integration structures. Once again, in the to-be-reported experiments, an ideal observer would simply ignore the probabilistic feedback and classify an object on the basis of its position with respect to a set of optimally placed, rule-based boundaries. By contrast, if exemplar-retrieval processes play a role even in situations involving rule-based structures, classification responding should be slower and more error prone for stimuli that receive probabilistic feedback.

Although evidence of exemplar influences in rule-based categories has previously been reported in some well-known studies (e.g., Allen & Brooks, 1991; Regehr & Brooks, 1993), the present experiments were designed to provide more incisive tests of that hypothesis. In particular, as we argue more fully in the online supplement to this article, the previous studies showed only that not all participants relied solely on certain minimally complex rules that may have been difficult to discover or use in the first place. Instead, the participants in those studies made use of information provided along additional stimulus dimensions that was advantageous for learning the categories. By contrast, in the present designs, the optimal rules should be easy to discover and use, and attending to the additional exemplar-specific information can only hurt performance.

EXPERIMENT 1

A potential way to get started in the present investigation would be to test a unidimensional rule-based structure by rotating the Figure 1 information-integration category structure 45º. However, exemplar models make allowance for the idea that participants attend selectively to relevant dimensions. Assuming that this form of selective attention takes place, there would be no psychological distinction between the probabilistic and deterministic pairs in Figure 1 (rotated 45º). Therefore, we needed to design alternative structures to contrast the exemplar and decision-boundary models.

The probabilistic rule-based category structure that we used in Experiment 1 is illustrated in Figure 2. The stimuli were composed of two highly separable dimensions (Garner, 1974; Shepard, 1964): the degree of saturation of a red rectangle and the left–right placement of a vertical line within the rectangle. We used separable-dimension stimuli in order to increase the plausibility that the participants might adopt logical rules as a basis for classification. There were four values per dimension, combined orthogonally to produce the 16-member stimulus set. As is illustrated in the figure, the 16 stimuli were divided into four categories, each described by a conjunctive rule. For example, an object is a member of Category A if it has a saturation value less than or equal to 2 and a line-position value greater than or equal to 3. The hypothesized rule-based boundaries for dividing the space into category regions are depicted in Figure 2. It is a rule-based structure because the boundaries are orthogonal to the coordinate axes. Indeed, similar multiple-category structures defined by conjunctive rules have been tested in a number of previous experiments designed to assess the properties of rule-based classification (e.g., Maddox, Filoteo, Heil, & Ing, 2004).

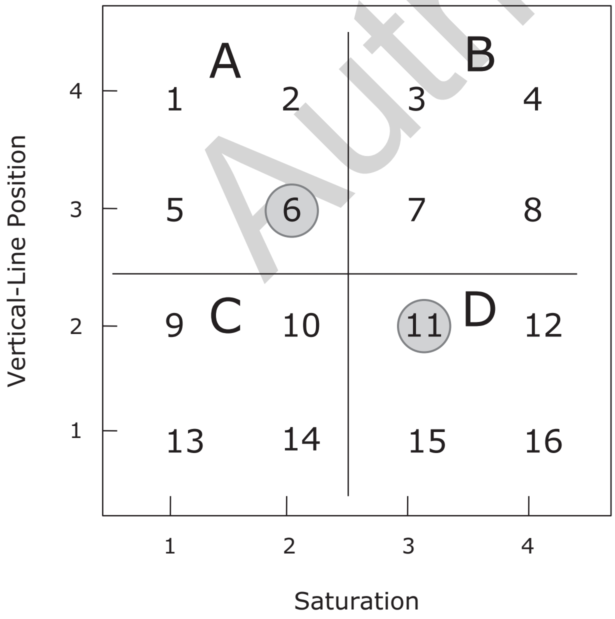

Figure 2. Schematic illustration of the conjunctive-rule category structures tested in Experiment 1. Each quadrant corresponds to a separate category. Across conditions, either Stimulus Pair 6–11 or Stimulus Pair 7–10 received probabilistic feedback. The horizontal and vertical boundaries are the hypothesized rule-based boundaries for dividing the space into category regions.

The critical manipulation, illustrated schematically in Figure 2, is that two stimuli were assigned probabilistic rather than deterministic feedback. In the Figure 2 example, Stimulus 6 is assigned to Category A on 75% of the trials in which it is presented but is assigned to Category D on 25% of the trials. Likewise, Stimulus 11 is assigned to Category D on 75% of the trials but is assigned to Category A on 25% of the trials. In comparison, whereas Stimuli 6 and 11 receive probabilistic feedback, the comparison stimuli, Stimuli 7 and 10, receive deterministic feedback. The probabilistic pair (Stimuli 6 and 11) and the deterministic pair (Stimuli 7 and 10) are the same distance from the rule-based boundaries, so decision-boundary models predict that they should have the same RTs and accuracies. By contrast, exemplar models predict that the probabilistic pair should be classified more slowly and less accurately than the deterministic pair.

An important property of the design is that there is no plausible way to adjust the form of the rule-based boundaries to improve classification performance on the probabilistic stimuli, because the feedback is chosen randomly on each individual trial.[1] Another advantage is that to achieve accurate performance, participants must attend to both dimensions that compose the stimuli. Therefore, the exemplar model cannot escape its prediction that the probabilistic pairs will be classified less efficiently than the deterministic pairs by positing that an irrelevant dimension has been ignored.

The design also allows one to test for generalization effects of the probabilistic feedback. According to the EBRW model, test items do not retrieve only their own memory traces. Instead, they may also retrieve exemplars to which they are similar. Therefore, for the same reasons described earlier, stimuli that are similar to the probabilistic pairs might be classified more slowly than are comparison stimuli that are similar to the corresponding deterministic pairs. For example, in Figure 2, Stimuli 2 and 5 from Category A might be classified more slowly and less accurately than the comparison stimuli, Stimuli 3 and 8 from Category B, because the former border a probabilistic stimulus, whereas the latter border a deterministic one. The extent to which the exemplar model does indeed predict such generalization effects, however, depends on its parameter settings, so we do not state it as a strong prediction from the model.[2]

A final goal of Experiment 1 was to address an idea advanced by Jones, Love, and Maddox (2006)—namely, that evidence for exemplar retrieval in classification may reflect only very short-term memory processes (see also Petrov & Anderson, 2005; Stewart, Brown, & Chater, 2002). For example, in the present kinds of designs, effects of probabilistic feedback may arise only when a stimulus is associated with incorrect feedback on trial n, and the same stimulus is then presented on the following trial $n+1$. To address the concern, we extended the design by also including blocks of test trials in which no feedback was presented. Continued effects of the probabilistic feedback during the test blocks would provide evidence of longer term effects of the manipulation.

Method

Participants. Twelve participants were recruited from the Indiana University community. Eleven participated in five 1-h sessions, one per day, whereas 1 completed four sessions. (The extended data collection at the individual-participant level was intended to allow for detailed quantitative modeling. However, for reasons described later, we focus here on the major qualitative pattern of results, rather than on quantitative fitting.) Each participant received $8 per session, plus a potential $3 bonus per session, depending on performance. All of the participants had normal or corrected-to-normal vision and claimed to have normal color vision.

Stimuli. The stimuli were 225×150 pixel rectangles displayed in red (Munsell hue 5R, brightness value 5) with a 10-pixel-wide black border and a 100×10 pixel black interior vertical line that extended from the lower left corner of the rectangle. There were 16 stimuli composed of all combinations of four levels of saturation (Munsell chromas 6, 8, 10, and 12) and four positions of the vertical line (30, 40, 50, and 60 pixels from the left corner). The colors were generated by converting them into RGB values with the Munsell Conversion Software Version 8 (http://wallkillcolor.com/). Multidimensional-scaling studies were used to verify that the psychological structure of the stimuli matched closely the schematic one shown in Figure 2. Details of the scaling study are provided in the online supplement.

Design and Procedure. The stimuli were divided into four categories with the structure depicted in Figure 2. The center four stimuli—Stimuli 6, 7, 10, and 11—were designated the critical stimuli. In Condition 7/10, Stimuli 7 and 10 were the probabilistic critical stimuli, and Stimuli 6 and 11 were the deterministic critical stimuli. These assignments were reversed in Condition 6/11. The probabilistic stimuli were assigned to their nominal category on 75% of the trials. On the remaining 25% of the trials, the probabilistic stimuli were assigned to the diagonally opposite category. All remaining stimuli received deterministic assignments for their respective categories. Half of the participants participated in Condition 7/10 and half in Condition 6/11.

The categories were associated with the "S," "D," "K," and "L" keys on the keyboard. The participants responded to Category S with the third (ring) finger of their left hand, Category D with the left-hand middle finger, Category K with the right-hand middle finger, and Category L with the right-hand third finger. The category-to-quadrant mappings were balanced across participants with the constraint that each hand made the response associated with one deterministic and one probabilistic category. The mappings were also constrained to ensure that one probabilistic category was assigned to the third finger of one hand and the other to the middle finger of the opposite hand. The same was true for the deterministic categories. These constraints leave 16 of 24 possible category-to-quadrant assignments; 12 were randomly selected for use in the experiment.

In the first session, the participants completed 850 training trials (trials with feedback), with a short break every 213 trials. In Sessions 2–5, the participants completed seven sets of trials, with each set consisting of one block of 64 training trials followed by one block of 64 test trials (trials without feedback). During these sessions, the participants were given a short break after every two sets of training and test blocks.

During the training blocks, the 4 center critical stimuli were presented on 50% of the trials. On the remaining trials, 1 of the remaining 12 stimuli was presented. Stimulus selection on each trial was random within these constraints. Thus, not including the first day, each of the critical stimuli was presented roughly 224 times during the training blocks, whereas each of the remaining stimuli was presented roughly 75 times. The increased presentation probability for the critical pairs was intended to increase the power of the experimental manipulation. Because of the symmetric arrangement of the critical stimuli, this increased presentation frequency does not affect the form of the optimal rule-based boundaries.

The blocks of 64 test trials were divided into four subblocks of 16 trials each. Within each subblock, each of the individual 16 stimuli was presented once in a random order. Thus, during the course of the experiment, each individual stimulus was presented 96 times during the test blocks.

On each training trial, a fixation cross was presented in the center of the screen for 0.5 sec. Then, a stimulus was presented and remained on the screen until a response was made. Following the response, feedback was presented on the screen for 1 sec ("Correct" or "Incorrect," or "Too Slow" if the RT was greater than 5 sec).[3] Following the feedback, there was a 0.5-sec blank interval before the next trial. Test trials were identical to training trials with the exception that no feedback was presented.

The participants were instructed to keep their fingers resting on the appropriate keys and to respond as quickly and accurately as possible. The participants were told that the task is a difficult one, and so perfect accuracy may not be possible, and that whether they received a bonus was determined by their overall performance.

Results

We excluded the first day of training from the analyses. In addition, for each individual participant–item combination, we excluded trials in which the RT was longer than 3 *SD*s above the mean or was shorter than 150 msec. The latter procedures led to excluding less than 2% of the data for each individual participant.

As was explained above, we refer to the four center stimuli as the probabilistic and deterministic critical stimuli. In addition (see Figure 2), we will refer to Stimuli 1, 4, 13, and 16 as the *corner* stimuli. When a corner stimulus comes from the same category as a probabilistic or deterministic critical stimulus, we will refer to it as a *probabilistic* or *deterministic corner stimulus*, respectively. For example, in Figure 2, Stimulus 1 is a probabilistic corner stimulus, whereas Stimulus 4 is a deterministic corner stimulus. We refer to all remaining stimuli as *adjacent* stimuli, because they are horizontally or vertically adjacent to a single critical stimulus. When an adjacent stimulus neighbors a probabilistic or a deterministic critical stimulus, we will refer to it as a *probabilistic* or *deterministic adjacent stimulus*, respectively.

The correct mean RTs and accuracies for the critical pairs, adjacent stimuli, and corner stimuli are displayed as a function of feedback (probabilistic vs. deterministic), separately for the training and test blocks, in Figure 3. (Because the pattern of results did not differ as a function of condition [7/10 vs. 6/11], the figure collapses over that variable.) For vs. probabilistic pairs, a correct response was defined according to an ideal observer. For example, for a participant in Condition 6/11, the correct response for Stimulus 6 was Category A, regardless of the feedback provided on a given trial.

As can be seen in Figure 3, in both the training and the test blocks, mean RTs for the probabilistic critical pairs were considerably longer than those for the deterministic critical pairs. In addition, in both the training and the test blocks, mean accuracies for the probabilistic critical pairs were lower than those for the deterministic critical pairs. The same pattern of results, for both RTs and accuracies during both training and test, tends to be observed for the adjacent and corner stimuli, although the magnitude of the effects is somewhat diminished. These latter results provide evidence for a generalized effect of the probabilistic feedback manipulation on performance. The overall

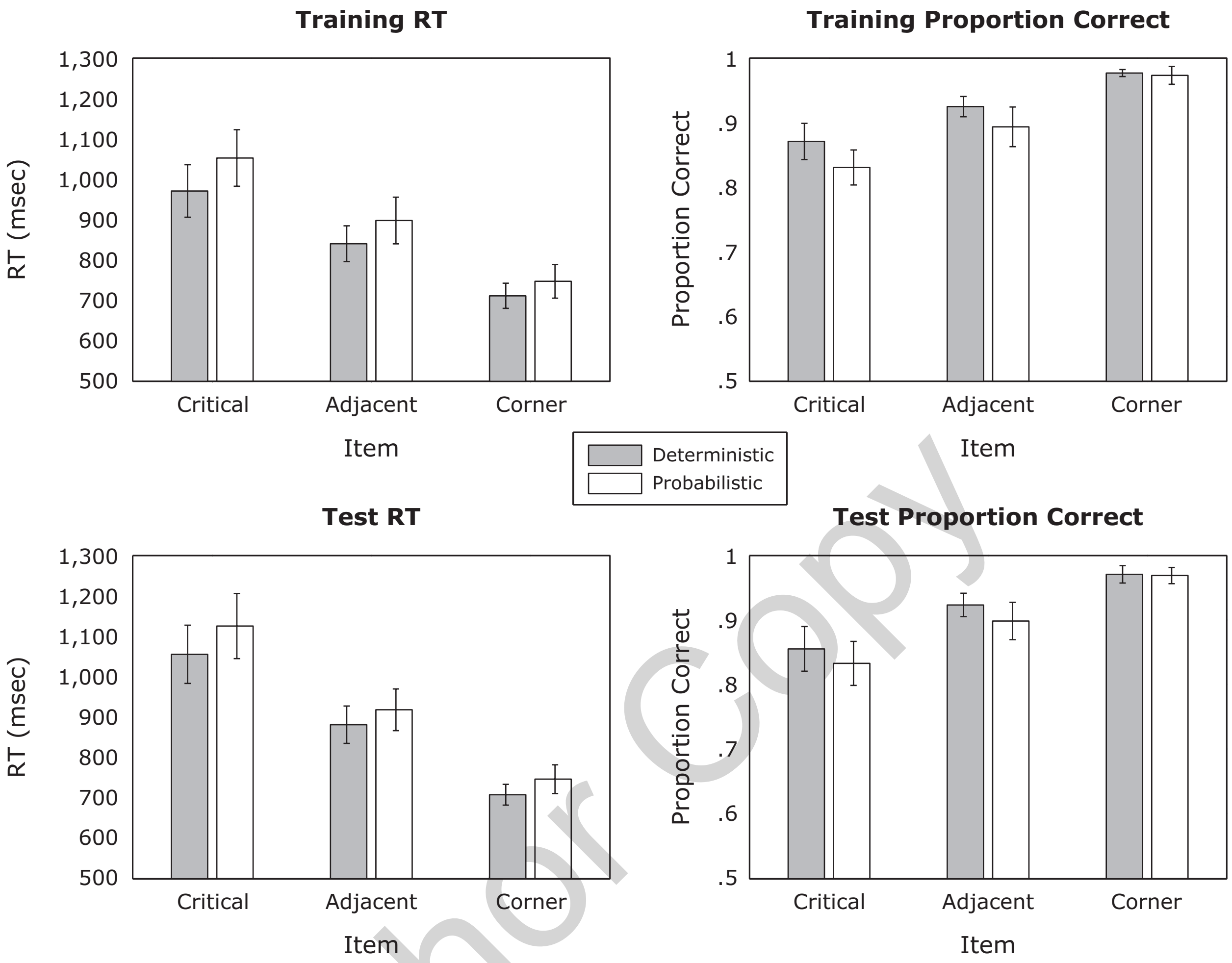

Figure 3. Mean correct response times (RTs) and accuracies in Experiment 1 plotted as a function of type of block (training vs. test), stimulus type (critical vs. adjacent vs. corner), and feedback (probabilistic vs. deterministic).

pattern of results is consistent with the predictions from the EBRW model and challenges the predictions from the rule-based distance-from-boundary models.

To confirm these observations, we conducted $2 \times 2 \times 3 \times 2$ ANOVAs of the data, with condition (7/10 vs. 6/11) as a between-subjects factor and with block type (training vs. test), item type (critical vs. adjacent vs. corner), and feedback (probabilistic vs. deterministic) as within-subjects factors. These analyses were conducted separately for the correct mean RTs and the accuracies.[4] We report here only the most important of the statistical test results. For the mean RTs, there was a main effect of feedback [$F(1,10) = 6.75$, $MS_e = 15{,}198.4$, $p = .027$], reflecting that stimuli associated with probabilistic feedback had longer RTs than did stimuli associated with deterministic feedback. In addition, the interaction between item type and feedback was significant [$F(2,20) = 4.62$, $MS_e = 1{,}051.0$, $p = .022$], reflecting that the slowdown for the probabilistic critical stimuli was more pronounced than for the probabilistic adjacent and corner stimuli. Likewise, for the accuracies, there was a main effect of feedback [$F(1,10) = 4.99$, $MS_e = 0.030$, $p = .049$], reflecting the lower accuracies associated with the probabilistic stimuli as compared with the deterministic ones. Again, the interaction between item type and feedback was significant [$F(2,20) = 3.88$, $MS_e = 0.010$, $p = .038$], although this particular interaction reflects, at least in part, a ceiling effect on the probability measure for the corner stimuli. Finally, the factor of block type (training vs. test) did not interact with any of the aforementioned effects, reflecting that the probabilistic feedback had relatively long-term effects on performance.

Although a full report goes beyond the scope of this article, we also conducted detailed individual-participants analyses of the complete RT distributions of the individual stimuli. Because of the assumption of a nonlinear relation between RT and distance, certain versions of the RT–distance model that posit increased perceptual variability for the probabilistic stimuli can predict lengthened mean RTs for those stimuli (see Nosofsky & Stanton, 2005, for details). However, these versions also predict that the very shortest RTs associated with the probabilistic critical stimuli will be shorter than the very shortest RTs associated with the deterministic ones. Our RT-distribution-based analyses revealed a pattern that went significantly in the opposite direction (a result predicted correctly by the EBRW model).

Discussion

In summary, the main pattern of results was that the probabilistic stimuli were classified more slowly and less accurately than were the corresponding deterministic stimuli that were the same distance from the decision boundary. There was also evidence of generalization effects, such that stimuli that were similar to the probabilistic pairs were classified more slowly and less accurately than were analogous stimuli that were similar to the deterministic pairs. This overall pattern of results is consistent with the predictions from exemplar models of classification but challenges the predictions from standard distance-from-boundary models. Furthermore, the results were observed in a domain involving salient rule-based category structures with stimuli composed of separable dimensions, conditions that might foster the application of decision-boundary strategies.

Nevertheless, a limitation of Experiment 1 was that the probabilistic and deterministic stimuli were each associated with separate category labels. For example, in Figure 2, the probabilistic stimulus, Stimulus 6, is associated with Category A, whereas the deterministic stimulus, Stimulus 7, is associated with Category B. Possibly, the probabilistic feedback manipulations led, at least in part, to category-level influences, rather than to exemplar-specific influences. For example, because of the probabilistic feedback associated with, say, Category A, an observer may have less confidence in making Category A responses, which is reflected in lengthened RTs and lower accuracies for stimuli associated with that category.

Although such influences are not formalized in standard distance-from-boundary models, it is straightforward to formalize them in extended decision-boundary models that posit stochastic forms of evidence accumulation (e.g., Ashby, 2000; Nosofsky & Stanton, 2005).

For example, Nosofsky and Stanton formalized a random-walk version of decision-boundary theory that implements distance-from-boundary effects. The representational assumptions in this random-walk distance-from-boundary (RW-DFB) model are the same as those in the standard version, with observers establishing decision boundaries to divide the stimulus space into category regions. However, instead of assuming a descriptive function for relating RTs to distance from boundary, a random-walk or counter process is assumed. Specifically, analogous to the EBRW model, there are counters initialized at 0 that take steps toward category criteria. On each step of the process, a percept is sampled at random from a distribution associated with the presented stimulus. If the percept falls in the Category A region defined by the decision boundaries, a counter associated with Category A is incremented by unit value. The perceptual sampling process continues until a criterion on one of the category counters has been reached, at which time the observer emits the categorization response.

Under suitable assumptions, it is straightforward to show that the RW-DFB model predicts the fundamental distance-from-boundary effect. However, a generalization of the model would allow different criterion settings across the alternative categories. In the case of the present Experiment 1, a reasonable hypothesis is that participants may set their criteria higher for categories that have been associated with probabilistic feedback, rather than with only deterministic feedback, thereby accounting for the lengthened RTs of the probabilistic stimuli.

It might still be possible to tell apart the EBRW model from this RW-DFB model on grounds of detailed quantitative fits to the individual-participant classification data. For example, for the Figure 2 design, the RW-DFB model tends to predict roughly equal effects of the probabilistic feedback on all members of a category. By contrast, the EBRW model predicts primarily exemplar-specific effects; however, depending on parameter settings, it predicts stimulus-generalization effects of varying magnitude as well. Especially because there are likely to be individual-participant differences, teasing apart these alternatives in the present design requires detailed, individual-participant model fitting. In this article, rather than pursuing this formal quantitative-comparison route, we decided instead to test a new category structure that yields sharply contrasting qualitative predictions from the models.

EXPERIMENT 2

The new category structure is illustrated schematically in Figure 4. As in Experiment 1, the stimuli varied along the dimensions of saturation and vertical-line position. As is illustrated in Figure 4, membership in Category A is defined by a conjunctive rule: A stimulus is a member of Category A if it exceeds a criterion value on the saturation dimension and exceeds a criterion value on the line-position dimension. Membership in Category B is defined by a complementary disjunctive rule: A stimulus is a member of Category B if it fails to exceed the criterion value on the saturation dimension or fails to exceed the criterion value on the line-position dimension. These rules are implemented in terms of decision boundaries that are orthogonal to the coordinate axes of the space.

Again, the key manipulation was that one pair of stimuli received probabilistic feedback, whereas a comparison pair received deterministic feedback. In the Figure 4 example, Stimulus 2 received Category A feedback on 75% of the trials but received Category B feedback on 25% of the trials, whereas Stimulus 5 received Category B feedback on 75% of the trials but Category A feedback on 25% of the trials. The comparison deterministic pair, analogously positioned in the stimulus space, was Pair 3–7. The assignment of probabilistic feedback and deterministic feedback to Pairs 2–5 and 3–7 was balanced across participants.

Again, the EBRW model predicts that the probabilistic pairs should be classified more slowly and less accurately, on average, than the deterministic pairs. By contrast, because the form of feedback does not affect the distance of the stimuli from the ideal-observer rule-based boundaries, distance-from-boundary models predict no

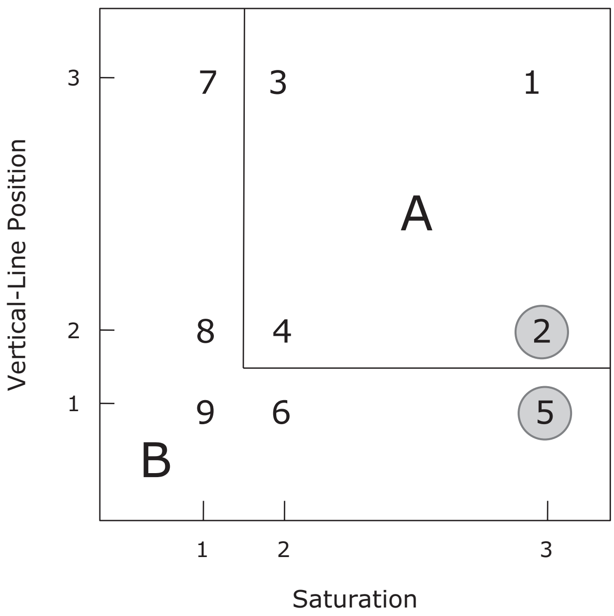

Figure 4. Schematic illustration of the category structure tested in Experiment 2. Category A is defined by a conjunctive rule, and Category B is defined by a complementary disjunctive rule. Across conditions, either Stimulus Pair 2–5 or Stimulus Pair 3–7 received probabilistic feedback.

effect of the probabilistic feedback. Furthermore, unlike in Experiment 1, the pairs that receive probabilistic versus deterministic feedback are not confounded with separate category labels. That is, one member of each pair is assigned to Category A, and the other member of each pair is assigned to Category B. Thus, the design sharply contrasts the predictions of the EBRW model with stochastic, random-walk versions of decision-boundary theory that allow separate criterion settings for different categories.

The Figure 4 structure was also designed to rule out another possible decision-boundary explanation of probabilistic-feedback effects. In particular, we considered the possibility that because the probabilistic stimuli from opposing categories differed from one another on a single dimension, participants might experience uncertainty in applying rules along that dimension. For example, in Figure 4, because of the inconsistent feedback provided for Pair 2–5, the participants might experience uncertainty in positioning the horizontal, rule-based boundary, leading to slowed decisions along that dimension. Crucially, however, if there is uncertainty in positioning the rule-based boundary, the influence of that uncertainty should extend to all stimulus pairs that are contrasted along that dimension. Thus, in Figure 4, any lengthening of RTs for Pair 2–5 should extend equally to Pair 4–6. Likewise, in the condition in which Pair 3–7 receives probabilistic feedback, any slowing of RTs for Pair 3–7 should extend equally to Pair 4–8. Under certain conditions, the EBRW model might predict similar response slowing due to stimulus-generalization effects from specific exemplars. However, as is illustrated in Figure 4, we arranged the stimulus spacings such that Pairs 2–5 and 4–6 were far separated and Pairs 3–7 and 4–8 were also far separated. Therefore, any effects of exemplar-specific generalization from the probabilistic critical stimuli should be minimized.

Finally, to ensure that any effects of probabilistic feedback assignments were relatively long term, the experiment also included test blocks in which all feedback was withheld.

In summary, for the present design, the EBRW model predicts lengthened RTs and lowered accuracies for the probabilistic critical pairs relative to the deterministic critical pairs, with far less generalized slowing for the adjacent pairs. By contrast, rule-based decision-boundary models predict either no effects of the probabilistic feedback or effects that extend equally to the probabilistic pairs and their adjacent, neighboring pairs in the space.

Method

Participants. The participants were 40 members of the Indiana University community. Each took part in a single session lasting approximately 45 min. Each participant received \$8 plus a possible \$3 bonus. Each reported having normal or corrected-to-normal visual acuity and normal color vision. There were 22 participants in Condition 2/5 and 18 in Condition 3/7.

Stimuli. The stimuli were the same type as those in Experiment 1. There were nine stimuli composed of all combinations of three levels of saturation (Munsell chroma levels 3.5, 5.5, and 18) and three positions of the vertical line (80, 64, and 20 pixels from the left-hand side of the rectangle). Multidimensional-scaling studies, described fully in the online supplement, were used to verify that the psychological structure of the stimuli matched closely the intended structure shown in Figure 4.

Design and Procedure. The stimuli were divided into two categories with the structure depicted in Figure 4. Stimulus Pairs 2–5 and 3–7 were designated the critical stimuli. In Condition 2/5, Stimuli 2 and 5 were the probabilistic critical stimuli, and Stimuli 3 and 7 were the deterministic critical stimuli. These assignments were reversed in Condition 3/7. On each individual trial of training, the probabilistic critical stimuli received feedback consistent with their nominal category with a probability of .75 and received feedback consistent with the contrast category with a probability of .25. All remaining stimuli received deterministic feedback assignments. The response keys for Categories A and B were always the "F" and "J" keys, respectively.

In the first part of the experiment, participants completed a 405-trial training phase (trials with feedback). During this phase, each of the nine stimuli was presented 45 times, in a random order. On each trial, a fixation cross was presented in the center of the screen for 0.5 sec. A stimulus was then presented in the center of the screen until a response was made. Following the response, feedback was presented on the screen for 1 sec ("Correct" or "Incorrect," or "Too Slow" if the RT was greater than 5 sec). Following the feedback, there was a 0.5-sec blank interval before the next trial. The participants received a break every 90 trials during the initial training phase and between the initial training phase and the final sets of training and test blocks.

Following the initial training, the participants were informed that their performance would be assessed to determine whether they would receive the bonus. The participants then completed a block of another 45 training trials, followed by a block of 45 test trials (trials without feedback), and then followed by another block of 45 training trials and another block of 45 test trials. Within each of these final blocks, each of the nine stimuli was presented five times in a random order. The procedure for the training blocks was the same as that in the initial training phase. The procedure for the test blocks was the same as that for the training blocks, except that no feedback was presented. The instructions were the same as those in Experiment 1.

Results

We analyzed the data from only the final four training and test blocks. We excluded from analysis those participants whose accuracy was less than 85% for the deterministic items in the final training and test blocks (because we judged that such participants might have been insufficiently motivated or might have failed to understand instructions). This procedure led to the removal of 7 participants from Condition 3/7 and 3 participants from Condition 2/5, leaving 15 participants in each condition. We excluded trials in which the RT was greater than 3 *SD*s above the mean for each participant–item combination or less than 150 msec. This procedure was applied separately for the training and test blocks and led to the removal of less than 1% of the trials.

Correct responses were again defined with respect to an ideal observer, not in terms of whether the response happened to match the probabilistic feedback provided on each individual trial. For example, in Figure 4, the correct response for Stimulus 2 is Category A, and the correct response for Stimulus 5 is Category B.

We will refer to the stimulus pair that is adjacent to the probabilistic critical pair as the *probabilistic adjacent pair*, whereas the pair that is adjacent to the deterministic critical pair is termed the *deterministic adjacent pair*. For example, in Figure 4, Pair 4–6 is the probabilistic adjacent pair and Pair 4–8 is the deterministic adjacent pair. Our analyses of the mean RTs and accuracies for the adjacent pairs include only the Category B members, because Stimulus 4 is common to each pair (see Figure 4). With respect to the adjacent pairs, the key question is whether performance on the probabilistic pair differs from that on the deterministic pair. To address this question, the analysis is restricted to the unique member of each pair.

The mean correct RTs and accuracies for the critical pairs and the adjacent pairs are displayed as a function of feedback (probabilistic vs. deterministic), separately for the training and test blocks, in Figure 5. (Again, because the pattern of results did not differ as a function of condition [2/5 vs. 3/7], the data in Figure 5 are collapsed over that variable.) As can be seen, in both the training and the test

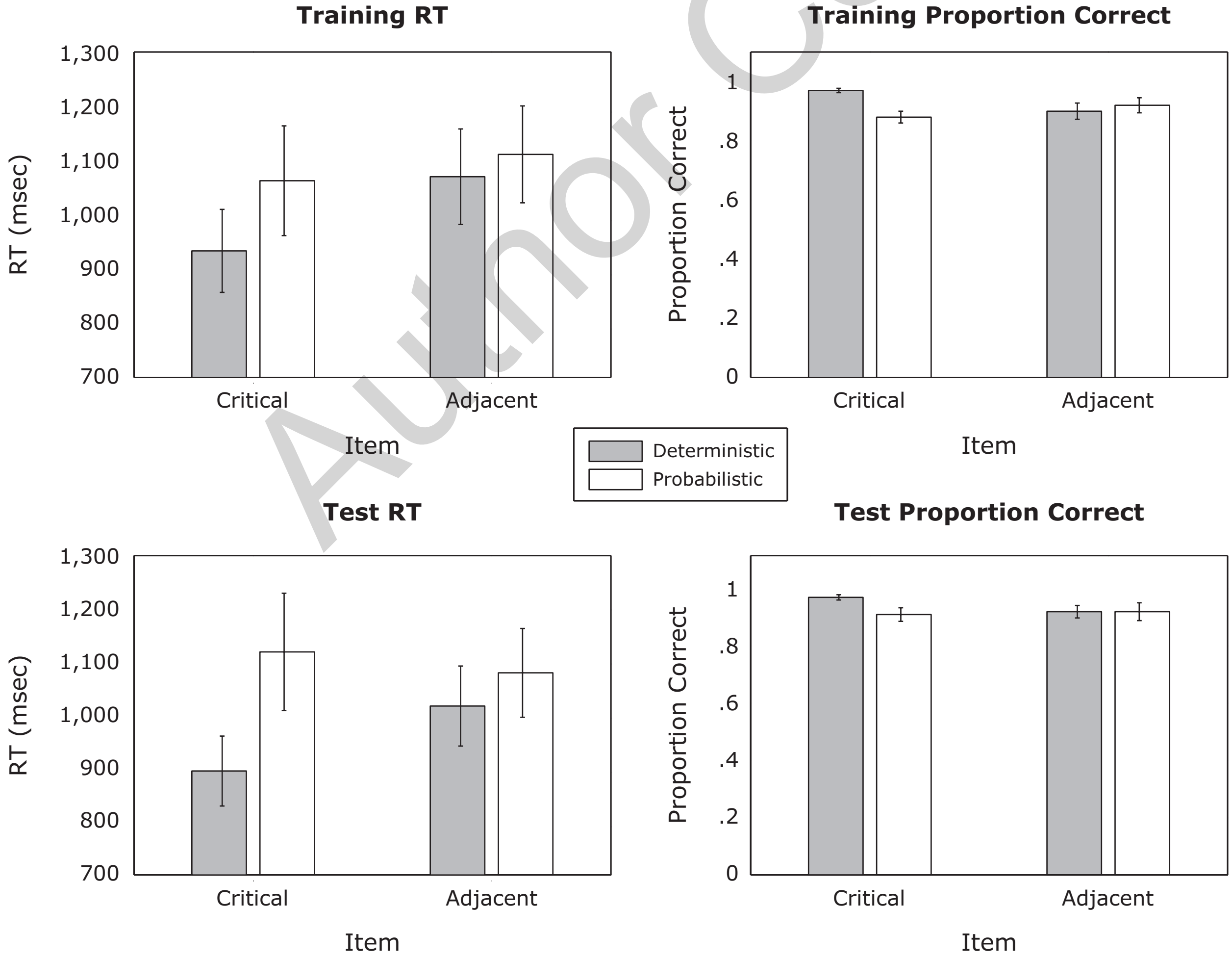

Figure 5. Mean correct response times (RTs) and accuracies in Experiment 2 plotted as a function of type of block (training vs. test), stimulus type (critical vs. adjacent), and feedback (probabilistic vs. deterministic).

blocks, mean RTs for the probabilistic critical pairs were considerably longer than those for the deterministic critical pairs. In addition, in both the training and test blocks, mean accuracies for the probabilistic critical pairs were lower than those for the deterministic critical pairs. By contrast, there were much smaller differences in performance between the probabilistic adjacent pairs and the deterministic adjacent pairs, in both the training and the test blocks. This pattern of results is consistent with the predictions from the EBRW model and challenges the predictions from the rule-based, distance-from-boundary models.

To confirm these observations, we conducted $2 \times 2 \times 2 \times 2$ ANOVAs of the data, with condition (2/5 vs. 3/7) as a between-subjects factor and block type (training vs. test), item type (critical pair vs. adjacent pair), and feedback (probabilistic vs. deterministic) as within-subjects factors. These analyses were conducted separately for the mean RTs and the accuracies. We report here only the most important of the statistical test results. For the mean RTs, there was a main effect of feedback [$F(1,28) = 11.68$, $MS_e = 73{,}585.62$, $p = .002$], reflecting that stimuli associated with probabilistic feedback had longer RTs than did stimuli associated with deterministic feedback. Crucially, the interaction between item type and feedback was statistically significant [$F(1,28) = 4.99$, $MS_e = 44{,}891.94$, $p = .034$], reflecting that the slowdown for the probabilistic critical stimuli was considerably more pronounced than for the probabilistic adjacent stimuli. Likewise, for the accuracies, there was a marginal main effect of feedback [$F(1,28) = 3.68$, $MS_e = 0.225$, $p = .07$], reflecting the lower accuracies associated with the probabilistic stimuli relative to the deterministic ones. Again, the interaction between item type and feedback was significant [$F(1,28) = 9.80$, $MS_e = 0.13$, $p < .01$], reflecting that the lowered accuracies for the probabilistic critical stimuli were much greater in magnitude than for the probabilistic adjacent stimuli. The interactions involving item type and feedback indicate that the effect of the probabilistic feedback was primarily an exemplar-specific one, extending mainly to the specific stimuli that received the probabilistic feedback. There was little if any generalization of the effect across the entire rule-based dimension. Finally, the factor of block type (training vs. test) did not interact with any of the aforementioned effects, reflecting that the probabilistic feedback had relatively long-term effects on performance.

In follow-up analyses, we attempted to ascertain whether some subset of our observers may have been rule-based responders. We reasoned that observers using deterministic rules and acting as ideal observers would be most likely to correctly classify the probabilistic stimuli into appropriate categories. Therefore, we focused consideration on those observers in each condition with the very lowest error rates and analyzed their RTs. Unfortunately, these follow-up analyses did not yield clear-cut conclusions. As was the case for the full set of observers, the subset of highest accuracy observers showed significantly longer RTs for the probabilistic critical pairs than for the deterministic critical pairs. Furthermore, the magnitude of the slowdown was reduced for the adjacent pairs relative to the critical pairs. Unfortunately, the interaction did not approach statistical significance. We cannot know whether the lack of a significant interaction reflects inadequate statistical power associated with the small subset of highest accuracy participants or if it is pointing to a group of rule-based responders who experienced uncertainty along the entire range of the probabilistic-feedback dimension. With larger sample sizes, the present design may be an excellent one for localizing rule-based versus exemplar-based responders.

GENERAL DISCUSSION

In summary, across two experiments and using rule-based category structures, we found that stimuli that received probabilistic feedback were classified more slowly and less accurately than were comparison stimuli that received deterministic feedback. There was also some evidence of generalization effects, in which RTs for stimuli that were highly similar to the probabilistic exemplars were lengthened. These effects were relatively long-term ones, extending into test blocks in which all feedback was withheld. The results are consistent with the predictions from exemplar models of classification, but challenge models that posit that classification RT is a function solely of the distance of a stimulus from rule-based decision boundaries. The results also challenged versions of rule-based decision-boundary models in which category evidence is accumulated in stochastic fashion and the required amount of evidence may differ for probabilistic and deterministic categories.

In our view, these results are particularly intriguing, given that the use of the simple rule-based decision boundaries would have been the optimal strategy for performing the tasks. To reiterate, for the present tasks, the best strategy would have been to ignore the probabilistic feedback associated with specific exemplars and to classify solely in accordance with the logical rules. Although Nosofsky and Stanton (2005) also found evidence for exemplar-retrieval effects in a similar design, those results were limited to a situation involving an information-integration category structure, in which no easily verbalizable rules were available. Therefore, the present results lend considerable generality to the previous findings and provide rather impressive evidence for a role of exemplar-retrieval processes in perceptual classification.

An interesting direction for future research would be to consider multiple-system accounts of performance in our tasks. According to the COVIS model of Ashby et al. (1998), there are two systems responsible for category learning: an explicit system that forms verbalizable rules and an implicit system that is mediated by procedural learning. A fundamental assumption in past applications of COVIS is that rule-based categories, such as the conjunctive-rule category structures tested in the present research, are learned by the explicit rule system. Indeed, in innumerable previous studies, COVIS theorists have modeled performance in such tasks by fitting rule-based decision-boundary models to the data. The results from the present research, however, challenge the assumption that performance in such tasks is mediated solely by application of these explicit rules. A COVIS theorist might argue that because of the inclusion of probabilistic stimuli in our designs, the participants' confidence in the rules was weakened enough for other processes to intrude. However, COVIS theorists have tested numerous previous designs involving rule-based category structures in which error rates were high because of overlapping category distributions or hard-to-discriminate stimuli. Despite the errors in responding, rule-based decision boundaries were presumed to govern performance. Therefore, a COVIS theorist would need to explain why other processes would intrude only in our present designs but not in previous ones that involved errors in responding.

An alternative multiple-system approach is Erickson and Kruschke's (1998) ATRIUM model. According to this model, a rule-based module and an exemplar-based module operate in concert, and the system learns which module to apply in different regions of the category space. In essence, ATRIUM subsumes exemplar models of classification as a special case, so arguments in favor of exemplar models must rely on parsimony. Nevertheless, to the extent that an ATRIUM account of performance in our tasks is correct, perhaps logical-rule processes predominated in those regions of our category structures that involved deterministic feedback assignments, whereas exemplar-retrieval processes operated in regions that involved probabilistic feedback. If so, perhaps brain-imaging techniques might reveal distinct neural signatures reflecting alternative classification strategies operating in probabilistic and deterministic regions of the space. We leave these and related questions motivated by our study as issues for future research.

AUTHOR NOTE

This work was supported by Grants FA9550-08-1-0486 from the Air Force Office of Scientific Research and MH48494 from the National Institute of Mental Health. Correspondence concerning this article should be addressed to R. M. Nosofsky, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405 (e-mail: nosofsky@indiana.edu).

REFERENCES

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, **120**, 3-19.

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, **44**, 310-329.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, **105**, 442-481.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, **55**, 11-27.

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

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of perceptual separability and perceptual integrality in speeded classification. *Journal of Mathematical Psychology*, **38**, 423-466.

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

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154-179.

Cohen, A. L., & Nosofsky, R. M. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, **47**, 150-165.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, **127**, 107-140.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, **407**, 630-633.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: An integration of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, **117**, 309-348.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, **32**, 108-154.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 316-332.

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.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, **124**, 161-180.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 695-711.

Lamberts, K. (2000). Information accumulation theory of categorization response times. *Psychological Review*, **107**, 227-260.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492-527.

Maddox, W. T., & Ashby, F. G. (1996). Perceptual separability, decisional separability, and the identification–speeded classification relationship. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 795-817.

Maddox, W. T., Filoteo, J. V., Heil, 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, & Cognition*, **30**, 227-245.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random-walk model of speeded classification. *Psychological Review*, **104**, 266-300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception & Performance*, **31**, 608-629.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review*, **112**, 383-416.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General*, **122**, 92-114.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, **133**, 63-82.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**, 54-87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 3-11.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, **2**, 442-459.

Verguts, T., Storms, G., & Tuerlinckx, F. (2003). Decision-bound theory and the influence of familiarity. *Psychonomic Bulletin & Review*, **10**, 141-148.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

NOTES

1. Technically, the optimal likelihood-based boundaries that would maximize accuracy have very slight curvature in the regions of the probabilistic critical stimuli. The differences in accuracy produced by the likelihood-based boundaries and the optimal rule-based boundaries are minuscule. Regardless, the present design is intended to contrast predictions from exemplar models and models that posit the formation of rule-based boundaries. It is not intended to contrast exemplar models with the general class of decision-boundary models. In our view, the hypothesis that observers would adopt rule-based boundaries for the present design is an extremely plausible one and merits careful investigation.

2. In the formal EBRW model, similarity is an exponential decay function of psychological distance (Shepard, 1987). The steepness of the function is governed by a sensitivity parameter. When sensitivity is high, the function is steep. In this case, test items tend to self-retrieve their own exemplar traces, and there is little in the way of stimulus-generalization effects. For lower settings of the sensitivity parameter, the similarity gradient broadens, and there is more in the way of stimulus-generalization effects.

3. Our intent in the experimental design was to provide the actual category assignment on each individual trial as part of the feedback. Unfortunately, we discovered only after the completion of the experiment that only "Correct" and "Incorrect" were provided as feedback. Thus, in the present four-category experiment, if a participant responded incorrectly, he or she was unable to determine with certainty the actual category assignment on that trial. This problem undoubtedly slowed down the rate of category learning. However, we do not include the first session of training in our analyses, which included 850 trials. Furthermore, the main qualitative predictions from the competing models are the same, regardless of whether the actual category assignment was indicated as part of the feedback. This issue will, however, cause complications in our goal of eventually quantitatively modeling the data, because there will be ambiguity in determining the exact proportion of exemplar traces associated with each of the category labels. Because of these complications, we do not report quantitative modeling in this article, but focus instead on the major qualitative patterns of results of interest.

4. With the goal of stabilizing variances, we transformed the probability-correct data (p) using the transform

$$p' = 2 \cdot \arcsin\left(\sqrt{p - 1/(2n)}\right),$$

where n is the number of observations on which the proportion correct is based (Winer, 1971, p. 400).

SUPPLEMENTAL MATERIALS

The results of our multidimensional-scaling studies, as well as a discussion of previous studies that found exemplar-based effects in rule-based category structures, may be downloaded from http://mc.psychonomic-journals.org/content/supplemental.

(Manuscript received October 31, 2009;
revision accepted for publication March 18, 2010.)

Online Supplement to Nosofsky & Little (2010)

This supplement to Nosofsky and Little (2010) has two parts. Part 1 provides more detailed arguments about the reason why the demonstrations of exemplar-retrieval processes in rule-based categories in our article go significantly beyond the well known previous studies of Allen and Brooks (1991) and Regehr and Brooks (1993). Part 2 provides details of the multidimensional scaling studies used to confirm that the psychological structures of the stimuli we used conform closely to the intended experimental designs.

Part 1: Comparison to Previous Studies

As noted in our Introduction, previous well known studies conducted by Allen and Brooks (1991) and Regehr and Brooks (1993) have also provided evidence for a role of exemplar-retrieval processes in contexts involving rule-based category structures. However, as we now explain, the present form of evidence goes beyond that earlier work in important ways. The basic category structure tested in those earlier studies is shown in supplementary Table S1. The stimuli varied along five binary-valued dimensions, with four training exemplars in Category A and four training exemplars in Category B. The "rule" determining category membership was a "2-of-3" rule defined over Dimensions 1-3: As can be seen in the table, if a stimulus has logical-value 1 on at least two of the first three dimensions, then it belongs to Category A; otherwise, it belongs to Category B. To assess the application of this logical rule, Allen and Brooks and Regehr and Brooks included some critical transfer stimuli in a test phase. A "positive match" (PM) item was one that satisfied the logical rule for a category and that was also highly similar to a specific training exemplar from that same category, because it matched that exemplar on its two "irrelevant" dimensions (4 and 5). For example, in Table S1, item PM1 is highly similar to training-exemplar A1, satisfying the Category-A logical rule and also matching training-exemplar A1 on Dimensions 4 and 5. By contrast, a "negative match" (NM) item was one that satisfied the logical-rule for a category, but did not have high similarity to any specific training exemplars from the category, because of mismatches on the two "irrelevant" dimensions. For example, in Table S1, item NM3 satisfies the Category-A logical rule, and matches training-exemplar A3 on the rule-defining dimensions, but it mismatches that training exemplar on both irrelevant dimensions. The fundamental result obtained by Allen and Brooks and Regehr and Brooks was that, under numerous experimental conditions, subjects classified the positive-match items with greater accuracy than the negative-match items, providing evidence for a role of "specific exemplars" in application of the rule.

The most important point we wish to make here, however, is that in the Table-S1 category structure, a wide variety of logical rules are available for classifying correctly the training exemplars. In many cases, evidence for a role of "specific exemplars" may simply have involved subjects' use of these alternative rules, rather than the 2-of-3 rule assumed in Allen and Brooks. For example, according to Nosofsky, Palmeri, and McKinley's (1994) rule-plus-exception model, observers often proceed by trying to establish near-perfect single-dimension rules, and then memorize exceptions to "patch" those rules (cf. Medin, Wattenmaker, & Michalski, 1987). In the Table-S1 case, an observer might learn that logical-value-1 on Dimension 1 signals Category A and that logical-value 0 on Dimension 1 signals Category B. The only exception to be learned is that the combination 0-1 on Dimensions 4 and 5 reverses the category assignments just described. The reader may verify that this simple single-dimension-plus-exception rule correctly classifies all of the training exemplars, and classifies all of the positive-match transfer items into the same categories defined by the 2-of-3 rule. However, it classifies two of the negative-match transfer items (NM4 and NM8) into the opposite categories. A wide variety of such rules exist, and tend to predict the same basic pattern as described above, namely, classification of the PM items into the category consistent with the 2-

of-3 rule, but classification of some of the NM items into the opposite category. To the extent that subjects adopted these types of alternative rules, the averaged data would therefore show more accurate classification of the PM items than the NM items, where "accuracy" is computed according to the category assignments defined by the 2-of-3 rule.

In our view, therefore, Allen and Brooks showed only that, in the context of their studies, many subjects did not rely solely on the "minimum-complexity" 2-of-3 rule that was available for their structure. Instead, many subjects made use of information from other dimensions that was potentially <u>advantageous</u> for correctly classifying objects into the categories. We note as well that minimum-complexity rules such as the 2-of-3 rule are not always easy to discover, especially when embedded in structures that afford alternative rule-based opportunities. Indeed, empirical investigations suggest that humans often develop rules with higher specificity than is needed (Medin et al., 1987; Nosofsky, Clark, & Shin, 1989). Thus, the scenarios that we sketched above are highly plausible ones. An impressive aspect of the Allen and Brooks (1991) and Regehr and Brooks (1993) studies is that the effects of "specific exemplars" were observed even in conditions in which subjects were provided with explicit instructions to use the 2-of-3 rule. For a wide variety of reasons, however, some subjects may not have followed the instructions, and may have induced alternative logical rules, such as the example we provided above. Thus, the findings from the explicit-instructions conditions do not negate our basic point.

By contrast, in the experiments reported in the present article, the minimum-complexity optimal rules seem transparent and easy to discover. We are unaware of other options for plausible alternative logical rules that subjects might use. Furthermore, rather than being advantageous, making use of the "exemplar-specific" exceptions to the rules can only hurt performance for our category structures. Therefore, for the reasons outlined above, our view is that the present studies provide an extremely important complement to some of the earlier studies that showed evidence for a role of specific exemplars in influencing performance in rule-based category structures.

Part 2: Multidimensional Scaling Studies

To confirm that the psychological structure of the stimuli matched reasonably well the intended structure of the experimental designs, we conducted similarity-scaling studies with new, independent groups of subjects from the Indiana University community. The stimuli were the same as described in the Methods sections of Experiments 1 and 2. There were 17 and 19 subjects who participated in scaling studies for the Experiment 1 and Experiment 2 stimuli, respectively. On each trial, a pair of stimuli would be presented on the screen, one in the upper-left quadrant and the other in the upper-right quadrant. Subjects rated the similarity (1="least similar", 8="most similar") between each presented pair. Subjects were instructed to try to use the full range of ratings. For the Experiment 1 scaling study, there was a total of 120 unique pairings of the 16 stimuli; each pair was presented 7 times, for a total of 840 trials. For the Experiment 2 scaling study, there was a total of 36 unique pairings of the 9 stimuli; each pair was presented 15 times, for a total of 540 trials. The order of presentation of the pairs of stimuli was completely randomized. Likewise, the left-right placement of each stimulus in each pair was chosen randomly on each trial. The experiment was self-paced, and subjects took brief breaks after every quarter of the total trials.

We computed the averaged similarity rating for each pair of stimuli and derived two-dimensional scaling solutions for the stimuli by fitting these averaged ratings. The models assumed a linear relation between the ratings and the distances between stimuli in the space. Distance was computed by assuming a city-block distance metric. We conducted computer searches for the MDS-coordinate parameters that minimized the sum-of-squared deviations between the predicted and observed ratings. (Extremely similar solutions were derived using alternative ordinal-scaling methods.) For Experiment 1, the two-dimensional-scaling solution accounted for 95.1% of the variance in the averaged ratings. For Experiment 2, the two-dimensional scaling solution accounted for 98.4% of the variance in the averaged ratings. The derived scaling solutions for the stimuli in each experiment are shown in Figure S1. Inspection of the figure confirms that the psychological structure of the stimuli matches extremely closely the intended structure of the experimental designs (compare with Figures 2 and 4 in the article).

Supplementary References

Allen, S.W., & Brooks, L.R. (1991). Specializing the operation of an explicit rule. Journal of Experimental Psychology: General, 120, 3-19.

Medin, D.L., Wattenmaker, W.D., & Michalski, R.S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. Cognitive Science, 11, 299-339.

Nosofsky, R.M., Clark, S.E., & Shin, H.J. (1989). Rules and exemplars in categorization, identification, and recognition. . Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 282-304.

Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53-79.

Regehr, G., & Brooks, L.R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. Journal of Experimental Psychology: General, 122, 92-114.

Table S1. Category Structure Tested by Allen and Brooks (1991).

Category A

Training Exemplars

	Dimension 1	2	3	4	5
A1	1	1	1	0	0
A2	1	0	1	1	1
A3	0	1	1	0	1
A4	1	1	0	1	0

Positive-Match Transfer Items

PM1	1	0	1	0	0
PM2	1	1	1	1	1

Negative-Match Transfer Items

NM3	0	1	1	1	0
NM4	1	1	0	0	1

Category B

Training Exemplars

	Dimension 1	2	3	4	5
B5	0	1	0	1	1
B6	0	0	0	0	0
B7	1	0	0	0	1
B8	0	0	1	1	0

Positive-Match Transfer Items

PM5	0	0	0	1	1
PM6	0	1	0	0	0

Negative-Match Transfer Items

NM7	1	0	0	1	0
NM8	0	0	1	0	1

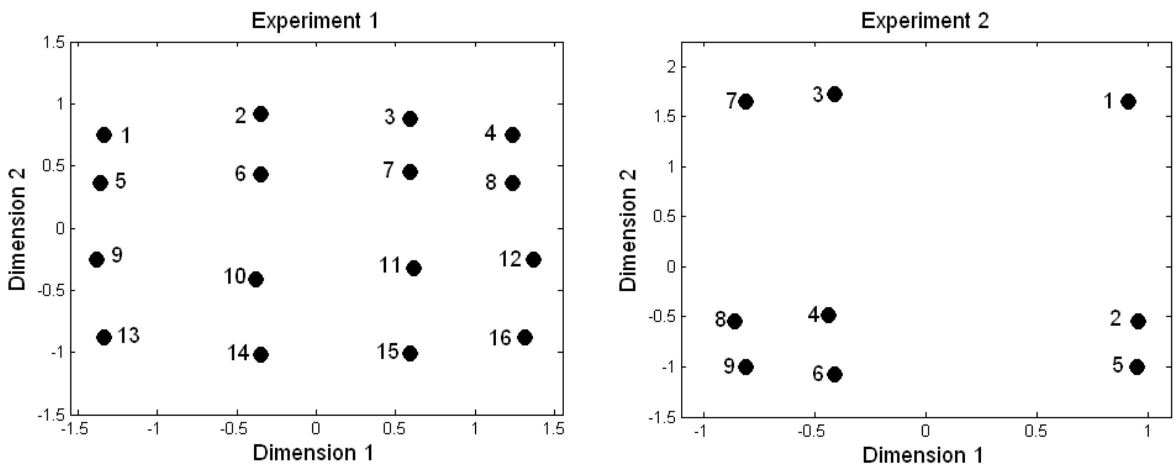
Experiment 1
Dimension 2
Dimension 1
Experiment 2
Dimension 2
Dimension 1

Figure S1.