Journal of Experimental Psychology:
Human Perception and Performance
2005, Vol. 31, No. 3, 608–629

0096-1523/05/$12.00 DOI: 10.1037/0096-1523.31.3.608

Speeded Classification in a Probabilistic Category Structure: Contrasting Exemplar-Retrieval, Decision-Boundary, and Prototype Models

Robert M. Nosofsky and Roger D. Stanton
Indiana University Bloomington

Speeded perceptual classification experiments were conducted to distinguish among the predictions of exemplar-retrieval, decision-boundary, and prototype models. The key manipulation was that across conditions, individual stimuli received either probabilistic or deterministic category feedback. Regardless of the probabilistic feedback, however, an ideal observer would always classify the stimuli by using an identical linear decision boundary. Subjects classified the probabilistic stimuli with lower accuracy and longer response times than they classified the deterministic stimuli. These results are in accord with the predictions of the exemplar model and challenge the predictions of the prototype and decision-boundary models.

A fundamental issue in the field of perceptual classification concerns the manner in which people represent categories in memory and the decision processes that they use for making classification judgments. Among the major formal models of perceptual classification are exemplar-retrieval, prototype, and decision-boundary models. According to exemplar-retrieval models (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986), people represent categories by storing individual exemplars of categories in memory, and they make classification decisions on the basis of the similarity of test items to these stored exemplars. According to prototype models (Posner & Keele, 1968; Reed, 1972; Smith, Murray, & Minda, 1997), a category representation consists of an idealized prototype, usually assumed to be the central tendency of the category training exemplars. And according to decision-boundary models (Ashby & Townsend, 1986), people use decision boundaries for dividing a multidimensional psychological space into category-response regions. These boundaries can correspond either to simple, verbalizable rules or to complex, nonverbalizable ones. Hybrid or multiple-system models have also been proposed that involve combinations of these types of representations and decision processes (Anderson & Betz, 2001; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994; Vandierendonck, 1995). However, the research reported in this article sought to develop contrasts among the predictions of the single-system models.

One of the emerging themes in the perceptual classification literature has been to test formal models not only on their ability to predict classification choice probabilities but on their ability to account for the actual time course of classification decision making (Anderson & Betz, 2001; Ashby, Boynton, & Lee, 1994; Ashby & Maddox, 1994; Cohen & Nosofsky, 2003; Lamberts, 1995, 1998, 2000; Maddox & Ashby, 1996; Nosofsky & Palmeri, 1997a, 1997b; Ratcliff & Rouder, 1998; Verguts, Storms, & Tuerlinckx, 2003). Thus, versions of the models have been developed that predict classification response times (RTs). We pursue this theme in the present article. Specifically, the purpose of this research was to conduct experiments to distinguish among the predictions of three formal models of choice probability and RT in tasks of speeded perceptual classification. The three models are representatives of the main model types described above: the exemplar-based random-walk (*exemplar-RW*) model (Nosofsky & Palmeri, 1997b), a newly proposed prototype-based random-walk (*prototype-RW*) model, and models based on the RT–distance hypothesis of decision-boundary theory (Ashby, 2000; Ashby et al., 1994).

Although the models are conceptually very different, they make surprisingly similar predictions across a variety of situations. In the present experiments, our key idea was to manipulate the probability with which specific exemplars were assigned to categories across different conditions of testing. As we show, the models make fundamentally different qualitative predictions in such a paradigm. Our goal was both to test these qualitative predictions and to evaluate the models on their ability to quantitatively fit the individual-subject accuracy and RT data.

It is important to note that Rouder and Ratcliff (2004) recently reported an extensive and highly systematic series of experiments for distinguishing between the exemplar-retrieval and decision-boundary models. Furthermore, as is the case in the present research, one of their key manipulations involved varying the probability with which individual stimuli were assigned to categories. We discuss the Rouder and Ratcliff experiments at length in this article. To anticipate, although the research themes are closely related, we suggest that our experimental manipulation provides an important qualitative contrast between the predictions of the competing models that was not present in Rouder and Ratcliff's designs. In addition, whereas Rouder and Ratcliff measured choice probability in the domain of unidimensional classification, we generalize the inquiry by examining both RTs and choice probabilities in tasks of speeded multidimensional classification. Thus,

Robert M. Nosofsky and Roger D. Stanton, Department of Psychology, Indiana University Bloomington.

This work was supported by Grant MH4848494 from the National Institute of Mental Health. We thank Jeff Rouder for criticisms and helpful suggestions for improving earlier versions of this article.

Correspondence concerning this article should be addressed to Robert M. Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405. E-mail: nosofsky@indiana.edu

our research provides a significant complement to the work of Rouder and Ratcliff.

We organize the remainder of the article as follows. First, we briefly review the formal exemplar-RW and decision-boundary models and also introduce the newly proposed prototype-RW model. Next, we explain the reason that the models tend to make similar predictions in various of the experimental paradigms that have been tested to date. We then review and evaluate some recent work that has attempted to distinguish the models. Finally, the main section of the article reports tests of the three models in a new experimental paradigm in which probabilistic assignments of exemplars to categories are manipulated.

Overview of the Formal Models

In this section, we describe the three formal models. In the to-be-reported experiments, the stimuli were Munsell colors of a constant hue that varied in their saturation and brightness. Such stimuli are classic examples of *integral-dimension* stimuli, in which the dimensions combine into relatively unanalyzable, unitary wholes (Garner, 1974; Shepard, 1987). Furthermore, in our experiments, the stimuli were assigned to one of two categories (A and B). Below, we describe the three formal models as they are applied in such a paradigm.

Exemplar-Based Random-Walk Model

According to the exemplar-RW model, people represent categories by storing individual exemplars in memory. Test items cause individual exemplars to be retrieved. The retrieved exemplars then drive a random-walk process (e.g., Busemeyer, 1985; Link, 1992; Luce, 1986; Ratcliff, 1978; Townsend & Ashby, 1983) that leads to classification decisions.

In the model, each exemplar is represented as a point in a multidimensional psychological space. Let x_{im} denote the value of exemplar i on psychological dimension m. When applied to the classification of integral-dimension stimuli, the distance between exemplars i and j is computed by using a weighted euclidean distance metric,

$$d_{ij} = (\Sigma\ w_m|x_{im} - x_{jm}|^2)^{1/2}, \quad (1)$$

where the w_ms ($0 \leq w_m \leq 1$, $\Sigma\ w_m = 1$) are free parameters representing the *attention weight* given to each dimension m. The similarity between exemplars i and j (s_{ij}) is an exponential decay function of psychological distance (Shepard, 1987), given by

$$s_{ij} = \exp(-c \cdot d_{ij}), \quad (2)$$

where c is an overall sensitivity parameter that describes the rate at which similarity declines with distance. The higher the value of c, the steeper the similarity gradient (i.e., the more discriminable are the exemplars in the psychological space).

Each exemplar resides in memory with strength M_j. In the baseline version of the model, the memory strengths are assumed to be proportional to the frequency with which each individual exemplar is presented in combination with given category feedback (Nosofsky, 1988b). When a test item is presented, it causes all exemplars to be activated. The activation for exemplar j, given presentation of item i, is given by

$$a_{ij} = M_j \cdot s_{ij}. \quad (3)$$

Thus, the exemplars that are most highly activated are those that have the greatest memory strengths and are highly similar to the test item.

When item i is presented, all category exemplars stored in memory race to be retrieved (cf. Logan, 1988). The race times are independent exponential random variables with rates proportional to the degree to which exemplar j is activated by item i (Bundesen, 1990; Logan, 1997; Marley, 1992; Marley & Colonius, 1992). Thus, the probability density that exemplar j completes its race at time t, given presentation of item i, is given by

$$f(t) = a_{ij} \cdot \exp(-a_{ij} \cdot t). \quad (4)$$

This assumption formalizes the idea that although the retrieval process is stochastic, the exemplars that tend to race most quickly are those that are most highly activated by the test item.

Finally, the exemplar that "wins" the race is retrieved and enters into a random-walk decision process. Specifically, the random-walk process is organized into a sequence of retrieval steps. In a two-category situation, the process operates 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 retrieval step. If x belongs to Category A, then the random-walk 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 categorization response is made. Otherwise, a new race is initiated, another exemplar is retrieved (possibly the same one as on the previous step), and the process continues.

Given the processing assumptions outlined above, Nosofsky and Palmeri (1997b) showed that on each step of the random walk, the probability (p_i) that the counter is increased in the direction of Category A is given by

$$p_i = \frac{S_{iA}}{(S_{iA} + S_{iB})}, \quad (5)$$

where S_{iA} denotes the summed activation of all currently stored Category-A exemplars given presentation of item i, and likewise for S_{iB}. (The probability that the counter is decreased in the direction of Category B is given by $q_i = 1 - p_i$.) So, for example, as the summed activation of Category-A exemplars increases, the probability of retrieving Category-A exemplars and thereby moving the counter in the direction of +A increases.

Given these random-walk processing assumptions, it is straightforward to derive analytic predictions of classification choice probabilities and mean RTs for each stimulus at any given stage of the learning process. The relevant equations are summarized by Nosofsky and Palmeri (1997b, pp. 269–270, 291–292). Because the current experiments used stimuli that varied along two dimensions and that were organized into two categories, the exemplar-RW model had six free parameters: the overall sensitivity parameter c; an attention-weight parameter w_1 (with $w_2 = 1 - w_1$); the random-walk criteria +A and −B; a scaling constant, k, for transforming the number of steps in the random walk into milliseconds; and a parameter, μ, representing the mean residual

time not related to classification decision making (e.g., encoding and response-execution time).[1]

One of the main predictions from the model is that the most rapid and accurate classification decisions should be made for those 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 will tend to come from the same category, so the random walk will march in consistent fashion to a single criterion. By contrast, items that are similar to exemplars from both categories should yield longer RTs. The reason is that the random-walk counter will tend to wander back and forth, sometimes retrieving exemplars from one category and other times retrieving exemplars from the contrast category.

It is important to note that in addition to yielding quantitative predictions of RTs, the exemplar-RW model provides a direct processing interpretation for the descriptive equations of choice probability found in the well-known *generalized context model* (GCM; Nosofsky, 1986; Nosofsky & Palmeri, 1997b, pp. 291–292). The GCM is an exemplar-based categorization model that has had a long record of success in predicting choice probabilities for individual stimuli in a wide variety of perceptual classification paradigms (e.g., McKinley & Nosofsky, 1995; Nosofsky, 1987, 1991; Nosofsky & Zaki, 2002). Specifically, consider a special case of the exemplar-RW model in which the criteria $+A$ and $-B$ are set an equal magnitude γ from the starting point of the random walk (i.e., $|A| = |-B| = \gamma$). In this case, the model predicts that the probability that item i is classified into Category A is given by

$$P(A|i) = \frac{{S_{iA}}^{\gamma}}{({S_{iA}}^{\gamma} + {S_{iB}}^{\gamma})}, \quad (6)$$

which is the GCM response rule (see Nosofsky & Palmeri, 1997b, p. 291). In this equation, S_{iA} and S_{iB} give the summed similarities of test item i to the exemplars of Categories A and B, respectively, whereas γ is a response-scaling parameter (Ashby & Maddox, 1993; McKinley & Nosofsky, 1995; Nosofsky & Zaki, 2002). When $\gamma = 1$, subjects respond by probability matching to the relative summed similarities of each category, whereas as γ grows greater than 1, subjects respond more deterministically with the category that yields the larger summed similarity. This role of the γ response-scaling parameter is discussed in greater detail in the *Applications to Past Experimental Data* section.

Prototype-Based Random-Walk Model

According to prototype models, people represent categories by forming abstract summary representations of categories, and they classify objects on the basis of their similarity to these prototypes. A prototype is usually assumed to correspond to the central tendency of a category's exemplars. Although exemplar and prototype models have been compared extensively on their ability to predict individual-stimulus choice probabilities in unspeeded classification paradigms, there has been little work comparing these models' predictions of speeded classification performance. In this section, we propose a prototype-RW model that is directly analogous to the exemplar-RW model, thereby allowing direct comparisons of RT predictions to be made.

In the prototype-RW model, the prototype of Category A is defined as the central tendency computed over Category A's training exemplars, and likewise for the prototype of Category B. The distance between a test item and the prototype is computed as in Equation 1; the similarity of the test item to the prototype is computed as in Equation 2; and the degree to which each prototype is activated and the rate at which it races is computed as in Equations 3 and 4. On each step of the random walk, the two prototypes race to be retrieved, and the winning prototype drives the random walk in the same manner as in the exemplar-RW model. Assuming that the prototypes have equal memory strengths, it is straightforward to show that on each step of the random walk, the probability of taking a step in the direction of Category A is given by

$$p_i = \frac{S_{iPA}}{(S_{iPA} + S_{iPB})}, \quad (7)$$

where S_{iPA} denotes the similarity of item i to the prototype of Category A. The mean RT and choice probability predictions of the prototype-RW model are then given by the same equations reported by Nosofsky and Palmeri (1997b, pp. 269–270) for the exemplar-RW model, with the exception of the new computation of p_i given in Equation 7 above.

A special case of interest arises when the random-walk criteria are set an equal magnitude γ from the starting point of the random walk. In this case, the prototype-RW model predicts that the probability with which test item i is classified into Category A is given by

$$P(A|i) = \frac{{S_{iPA}}^{\gamma}}{({S_{iPA}}^{\gamma} + {S_{iPB}}^{\gamma})}, \quad (8)$$

where γ is the response-scaling parameter. Equation 8 has been used extensively in previous work in applying prototype theory to the prediction of choice probabilities. However, as explained in previous work (e.g., Ashby & Maddox, 1993; Nosofsky & Zaki, 2002), if one limits consideration to the prediction of choice probabilities, then in the prototype model, the γ response-scaling parameter cannot be estimated separately from the overall sensitivity parameter c, so it is typically held fixed at 1. However, in the present RT domain, the value of γ (i.e., the values of $+A$ and $|-B|$) cannot be held fixed at 1 if the prototype-RW model is to provide plausible predictions of speeded classification performance. In such a case, for example, the model would predict that all stimuli are classified with equal response speed, regardless of their difficulty.[2] The parameters in the prototype-RW model are the same as in the exemplar-RW model: overall sensitivity parameter c, attention-weight parameter w_1, random-walk criteria $+A$ and $-B$, scaling constant k, and residual-time parameter μ.

[1] The version of the exemplar-RW model tested in this study differs from the original version in some minor respects. First, in the original version, the time to take each individual step in the random walk (T_{step}) was given by $\alpha + t$, where α is a constant term associated with each step, and t is the time to retrieve the winning exemplar. Because the stochastic retrieval-time component does not add materially to the steady-state predictions from the model, for simplicity, we now instead set T_{step} equal to unit value (see also Cohen & Nosofsky, 2003). In addition, some previous applications of the exemplar-RW model have included a background-noise parameter, representing the rate at which background elements stored in memory race against the stored exemplars to enter into the random walk. Because the background-noise parameter is important mainly for modeling initial learning, it is not included in the present model fits.

[2] Specifically, in the exemplar-RW and prototype-RW models, RT is determined by the total number of retrieval steps required to complete the random walk. If $\gamma = 1$, then the random walk is always completed in a single step, regardless of stimulus difficulty.

Decision-Boundary Model

According to decision-boundary theory (Ashby & Townsend, 1986), people use decision boundaries for dividing a perceptual space into category-response regions. Test items are assumed to give rise to noisy representations in the multidimensional perceptual space. For simplicity, in this article, we assume that the perceptual representations are independently and normally distributed along each dimension, with variance σ_p^2. (We consider the implications of some more complex assumptions as well.) If a test item gives rise to a point in Region A of the space, then the observer responds with Category A.

In most applications of decision-boundary theory, it is assumed that the observer uses a decision boundary that is optimal in form (i.e., a decision boundary with a functional form that would maximize the observer's proportion of correct classifications; Maddox & Ashby, 1993). In the to-be-reported experiments, the optimal decision boundary is linear in form, regardless of the probabilistic assignments of exemplars to categories. Thus, in this article, we focus on the predictions of the linear decision-boundary model.

Past approaches to generating RT predictions from decision-boundary theory involved application of the RT–distance hypothesis (Ashby et al., 1994). According to this hypothesis, mean RT is a decreasing function of the distance of a stimulus from the decision boundary. To generate quantitative predictions, specific assumptions are needed of the function relating RT to distance-from-boundary. In past tests, Maddox and Ashby (1996) found strongest support for an exponential function in which mean decision time (MDT) is given by

$$MDT = k \cdot \exp(-\beta \cdot D), \tag{9}$$

where D is distance-from-boundary, β determines the rate at which RT decreases with distance, and k is a scaling parameter. We assume this exponential model in deriving the quantitative predictions from decision-bound theory. It is important, however, to note that regardless of the specific quantitative function that is assumed, the linear decision-boundary model makes the same fundamental qualitative predictions of the effects of our probabilistic assignments of exemplars to categories.

In the present applications, the linear decision-boundary model uses six free parameters: a slope (m) and y-intercept (b) of the best-fitting linear decision boundary, the perceptual-variance parameter σ_p^2, the rate parameter β, the scaling constant k, and the mean residual-time parameter μ.

For completeness, in Appendix A, we also describe a random-walk version of the linear decision-boundary model (for a similar development using a continuous-time diffusion process, see Ashby, 2000). This random-walk version of the linear decision-boundary model has the same form as the exemplar-RW and prototype-RW models, except that the step probabilities are now determined by distance-from-boundary rather than by the retrieval of exemplars or prototypes. The random-walk version of the linear decision-boundary model yielded slightly better fits to our speeded classification data than did the standard RT–distance version, but it did so at the expense of an additional free parameter. Because none of our conclusions are changed by this, in this article we report the fits of only the standard RT–distance version.

Applications to Past Experimental Data

In early tests of classification RT predictions, Ashby et al. (1994) conducted experiments in which subjects classified objects from two bivariate, normally distributed categories. An illustration of their paradigm is shown in Figure 1, in which the category distributions A and B have the same variance along each of their dimensions. On each trial, a stimulus is selected randomly from one of the two categories, the subject classifies it as rapidly as possible, and the correct category label is then provided by the experimenter. Note that because the category distributions are overlapping, it is impossible to achieve perfect accuracy in such a paradigm.

The diagonal line in Figure 1B is the optimal decision boundary for separating the categories. An ideal observer will maximize his or her proportion of correct responses by classifying all items to the upper left of the boundary into Category A and all items to the lower right into Category B. Ashby and colleagues have observed that in this type of paradigm, individual subjects often make classification responses in a near-deterministic fashion in accordance with such an optimal decision boundary (Ashby & Gott, 1988; Ashby & Maddox, 1992). Furthermore, in the speeded classification version of this task, Ashby et al. (1994) found strong support for the RT–distance hypothesis. They observed a strong

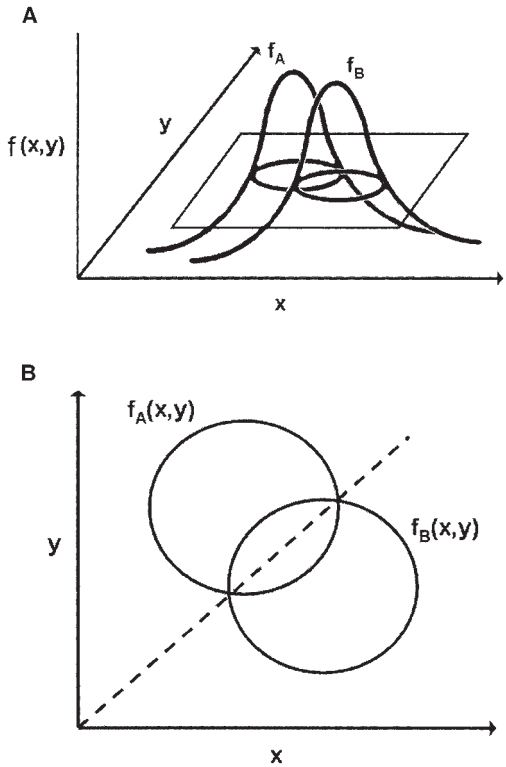


Figure 1. A: Schematic illustration of an experimental paradigm (see Ashby, Boynton, & Lee, 1994) in which subjects classify objects into two bivariate normal category distributions. f_A denotes the bivariate density associated with category distribution A, and f_B denotes the bivariate density associated with category distribution B. B: Equal-probability cross-sections of the bivariate normal distributions projected onto the x–y plane. The dashed diagonal line is the optimal decision boundary for dividing the x–y plane into response regions.

negative correlation between RTs and distance-from-boundary such that stimuli far from the boundary tended to be classified more rapidly than were stimuli close to the boundary.

Note that the γ response-scaling parameter in the GCM is crucial for allowing that model to account for the pattern of near-deterministic responding in this paradigm. Recall that with $\gamma = 1$ in Equation 6, the GCM predicts that subjects will respond by probability matching to the relative summed similarities of each category. Maddox and Ashby (1993) provided clear evidence that individual subjects responded more deterministically than predicted by this probability-matching rule. McKinley and Nosofsky (1995) showed that with γ allowed to vary freely, the GCM provided quantitative accounts of accuracy data in this paradigm that were as good as those provided by decision-boundary theory. And because the exemplar-RW model provides a direct process-model interpretation of the emergence of the γ response-scaling parameter, it accounts for such data as well.

Furthermore, Nosofsky and Palmeri (1997b) conducted simulations demonstrating that the exemplar-RW model successfully accounted for the RT data reported by Ashby et al. (1994; for details, see Nosofsky & Palmeri, 1997b, pp. 272–273). In general, in the paradigm illustrated in Figure 1, an exemplar that is far from the boundary tends to be highly similar only to exemplars from its own category. Thus, on each step of the random walk, exemplars from the correct category are retrieved, and the counter marches in consistent fashion to the appropriate category criterion. By contrast, an exemplar that lies close to the boundary tends to be similar both to exemplars from its own category and to exemplars from the contrast category. Thus, the random walk wanders back and forth, and decision times are longer.

In addition to considering performance in paradigms involving bivariate normal categories, Nosofsky and Palmeri (1997b, Experiment 1) tested the exemplar-RW model and the decision-boundary model in designs involving a smaller number of stimuli, with each individual stimulus presented on multiple trials. In such designs, one can measure choice probabilities and RTs for individual stimuli and provide rigorous tests of the models' ability to quantitatively fit the individual-stimulus data. Despite their vast conceptual differences, the quantitative fits provided by the exemplar-RW model and the decision-boundary model were essentially the same, and the models could not be sharply distinguished (for details, see Nosofsky & Palmeri, 1997b, pp. 276–280).

The reason that the exemplar-RW and decision-boundary models make similar predictions is that distance-from-boundary and relative summed similarity tend to be highly correlated in such designs. As explained above, items that are far from the boundary tend to be highly similar to exemplars from their own category and not similar to exemplars from the contrast category.

The key to distinguishing between the predictions from the models is to develop paradigms in which distance-from-boundary and relative summed similarity are decoupled. In some past work, one approach to achieving this aim has been to manipulate the absolute frequency with which individual stimuli are experienced during classification training (Nosofsky & Palmeri, 1997b, Experiment 2; Verguts et al., 2003). The exemplar-RW model predicts that, all other things being equal, familiar stimuli should be classified more rapidly than unfamiliar ones, because increasing the frequency of an item boosts its summed similarity to the target-category exemplars. This prediction from the exemplar-RW has been confirmed in studies in which absolute frequency was manipulated experimentally across conditions. However, it is possible that effects of absolute frequency may involve "surprise" effects, and their locus may reside in psychological factors not associated with classification decision making. It is important, therefore, to seek converging evidence for such effects by using alternative experimental manipulations.

Rouder and Ratcliff (2004)

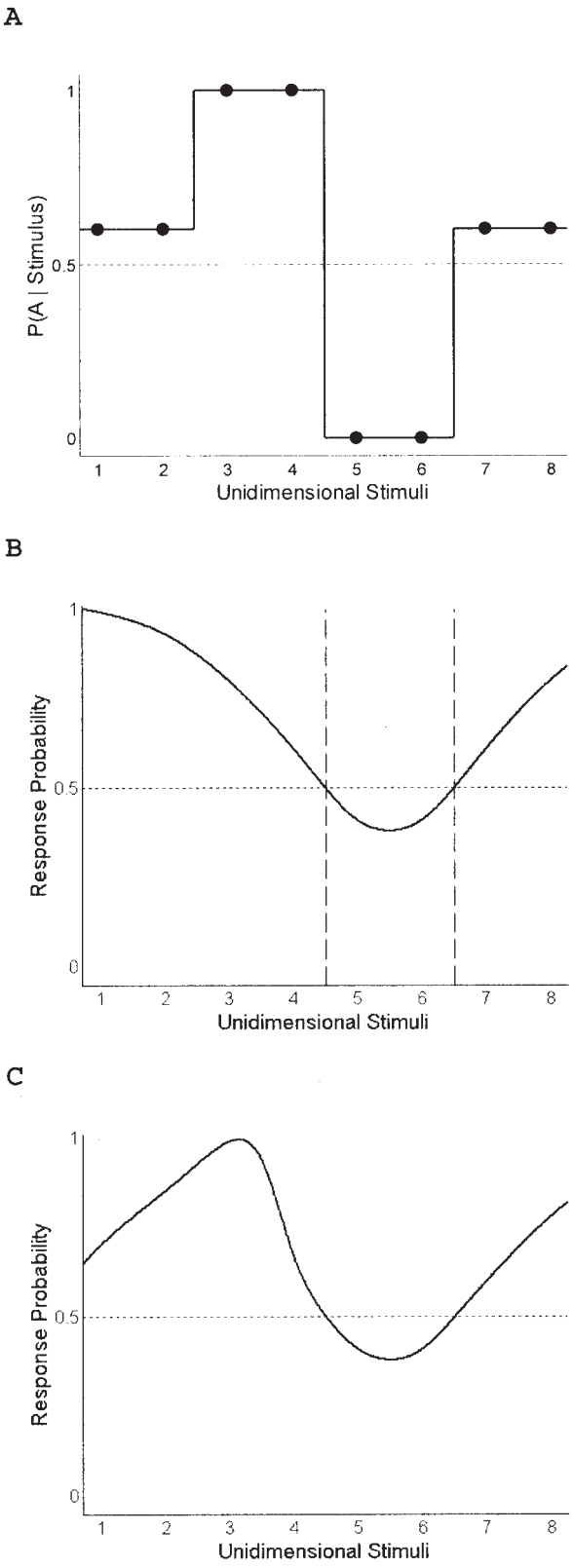
One such manipulation was carried out by Rouder and Ratcliff (2004) in a recent series of experiments involving unidimensional classification. The key idea in their experiments was to decouple distance-from-boundary and relative summed similarity by manipulating the probability with which individual stimuli were assigned to alternative categories. The design of a representative experiment from their studies is illustrated in Figure 2A. As illustrated in the figure, there were eight equally spaced stimuli varying along a unidimensional continuum. The stimuli were assigned to one of two categories (A and B). The middle stimuli were assigned deterministically to their respective categories. Thus, Stimuli 3 and 4 received Category-A feedback with probability 1, whereas Stimuli 5 and 6 received Category-A feedback with probability 0. By contrast, the extreme stimuli (Stimuli 1 and 2 and Stimuli 7 and 8) were assigned probabilistically to the categories. In the present illustration, all of the extreme stimuli received Category-A feedback with probability .60.

To apply decision-boundary theory, Rouder and Ratcliff (2004) assumed that subjects would partition the perceptual space by establishing cutoffs between Stimuli 4 and 5 and between Stimuli 6 and 7 (see Figure 2B). Any percept falling within the interior region defined by these cutoffs would be classified in Category B, whereas percepts falling outside these cutoffs would be classified in Category A.

Rouder and Ratcliff's (2004) design decouples stimulus probability and distance-from-boundary by placing them in opposition to one another. That is, Stimuli 1 and 2 are farther from the decision boundary than are Stimuli 3 and 4, but they receive Category-A feedback with lower probability. As a result, exemplar-retrieval and decision-boundary models tend to make contrasting predictions. The typical response-probability predictions from the decision-boundary model are as illustrated in Figure 2B: The farther away that an A stimulus is from the nearest cutoff, the higher should its Category-A response probability be. By contrast, the typical response-probability predictions from the exemplar model are as illustrated in Figure 2C: Because the exemplar model's predictions are influenced by the category-assignment probabilities, it tends to predict lower Category-A response probabilities for the extreme stimuli (Stimuli 1 and 2) than for the middle ones (Stimuli 3 and 4).

Rouder and Ratcliff's (2004) design does indeed place severe constraints on the predictions from the alternative models. Furthermore, these researchers conducted extensive and painstaking quantitative model-fitting analyses to determine the experimental conditions that tended to favor one model over the other. Their general pattern of observed results was that in conditions involving highly confusable stimuli in which it was difficult to discriminate among individual exemplars, the quantitative predictions favored the decision-boundary model over the exemplar model. By contrast, in conditions involving more discriminable stimuli, the quantitative predictions from the exemplar model were favored.

Despite this systematic pattern of observed results, the key point that we make here is to emphasize that in Rouder and Ratcliff's



(2004) design, the variables of distance-from-boundary and stimulus probability are pitted against one another, not manipulated as orthogonal experimental factors. Furthermore, the exemplar-retrieval model predicts that classification choice probabilities and RTs should be sensitive to both factors, with the relative impact of each factor depending on specific parameter settings and detailed assumptions in the modeling.

Indeed, in Figure 3 we illustrate predictions from a version of the exemplar model that does not conform to the typical pattern described by Rouder and Ratcliff (2004). Details of this modeling illustration are provided in Appendix B. In brief, this version of the exemplar model makes allowance for the reasonable idea that in situations involving highly confusable stimuli, one needs to model explicitly the sensory and memory noise associated with the stored exemplars (Nosofsky, 1988a, 1997). As can be seen in the figure, when allowance is made for the role of sensory and memory noise, the exemplar model can predict a response-probability gradient that increases monotonically with distance from the decision boundary, despite the probabilistic feedback associated with the extreme stimuli. Indeed, the predicted gradient matches the typical pattern that is predicted by the decision-boundary model extremely well.

We emphasize that the point of this illustration is not to claim that the exemplar-retrieval model is sufficient to account for all of Rouder and Ratcliff's (2004) data. It remains an open question, for example, whether an exemplar model that makes allowance for sensory–memory noise can quantitatively fit the data from their conditions involving highly confusable stimuli. Rather, we are suggesting only that in Rouder and Ratcliff's design, the qualitative contrast between the models may not be quite as sharp as is illustrated by the differing response-probability gradients in Figures 2B and 2C. Accordingly, there is a need to rely on quantitative fit indexes as a basis for comparing the models. As is well known, however, such indexes can be highly influenced by detailed formal assumptions that are not central to the key conceptual underpinnings of models. Furthermore, the quantitative fits that are achieved will also be influenced by the inherent flexibility (or complexity) of the competing models (e.g., Pitt, Myung, & Zhang, 2002).

Thus, although Rouder and Ratcliff's (2004) design places severe constraints on the alternative models, our view is that other approaches to developing qualitative contrasts would also be valuable. In the present experiments, we pursued the general tack taken by Rouder and Ratcliff, except we did not pit stimulus probability and distance-from-boundary against one another. Instead, we attempted to manipulate stimulus probability as an independent experimental factor while holding distance-from-boundary roughly constant. As we show, despite manipulating probabilistic categorization assignments across conditions, our design ensured that the optimal decision boundary, as well as the distance of individual stimuli to the boundary, remained unchanged across the

Figure 2. A: Schematic design of a representative experiment from Rouder and Ratcliff (2004). B: Typical predictions from the decision-boundary model. The vertical dashed lines represent cutoffs by which subjects partitioned perceptual space. C: Typical predictions from the exemplar-retrieval model. From "Comparing Categorization Models," by J. N. Rouder & R. Ratcliff, 2004, *Journal of Experimental Psychology: General, 133,* p. 65. Copyright 2004 by the American Psychological Association. Adapted with permission.

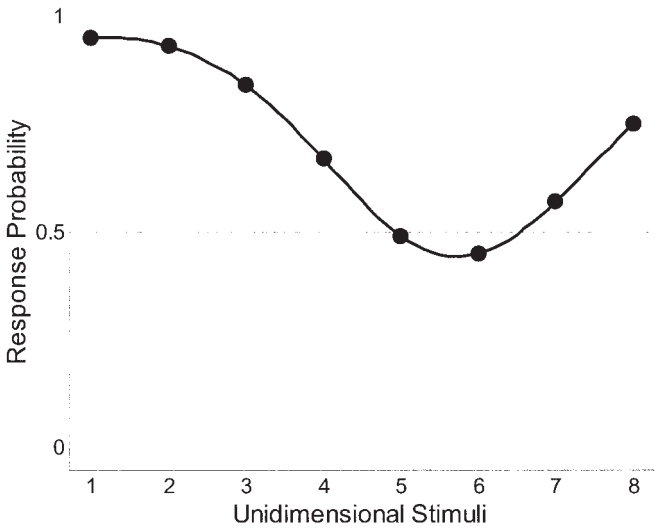


Figure 3. Illustrative predictions of performance in the Figure 2A design from a version of the exemplar model with sensory/memory noise (see Appendix B for details).

conditions. As a result, the exemplar-RW and decision-boundary models made sharply different qualitative predictions.

Our experiments differed from those of Rouder and Ratcliff (2004) in other important ways as well. First, recall that Rouder and Ratcliff examined choice probability in tasks of unidimensional classification, whereas our experiments examined both RTs and choice probability in tasks of multidimensional classification. The distinction between unidimensional and multidimensional classification is of major potential importance. In the domain of unidimensional classification, forming decision boundaries amounts to setting cutoffs (i.e., establishing single points) along a dimension. A similar psychological process operates in 2-D classification only when the decision boundaries are straight lines that are orthogonal to the coordinate dimensions. In this case, the observer establishes cutoffs along a single dimension while ignoring values along the second dimension. By contrast, in our experiments the presumed decision boundary was an oblique line that required integration of perceptual information from both dimensions. As noted by Ashby et al. (1998), this type of decision boundary is extremely difficult to verbalize, in contrast to what is involved in setting cutoffs along individual dimensions. Indeed, Ashby and colleagues have argued that the distinction between these types of decision boundaries is so fundamental that separate cognitive systems underlie their use (for an extensive discussion and review, see Ashby & Casale, 2003). It is critical, therefore, to test whether the types of results observed by Rouder and Ratcliff generalize to multidimensional domains.

Finally, the motivating theme of our research was to contrast the predictions of models of the time course of classification. Thus, we extended Rouder and Ratcliff's investigations by examining the effects of probabilistic exemplar assignments on RTs in addition to choice probabilities.

Experiment 1

The design of Experiment 1 is illustrated in Figure 4. The stimuli were 12 Munsell colors of a constant hue, varying in brightness and saturation. The colors were assigned to one of two categories (A and B). As illustrated in the figure, Colors 1–6 belonged to Category A, whereas Colors 7–12 belonged to Category B. Given our above-discussed simplifying assumptions about perceptual noise, the optimal boundary for separating the two classes of colors into response regions was the diagonal linear decision boundary illustrated in the figure. (We consider some alternative perceptual-noise assumptions in the *Theoretical Analysis* section of Experiment 2.)

The key experimental manipulation was that across conditions, either Stimulus Pair 4/8 or Stimulus Pair 5/9 received probabilistic feedback, whereas all other stimuli received deterministic feedback. Specifically, in Condition 4/8, Stimulus 4 received Category-A feedback on .75 of the trials, and it received Category-B feedback on .25 of the trials. Likewise, Stimulus 8 received Category-B feedback on .75 of the trials, and it received Category-A feedback on .25 of the trials. Analogous probabilistic feedback was assigned to Stimulus Pair 5/9 in Condition 5/9. We refer to these four centrally located stimuli (Pairs 4/8 and 5/9), which received probabilistic feedback across conditions, as the *critical stimuli*. The pair that received probabilistic feedback is the *probabilistic* critical pair, whereas the pair that received deterministic feedback is the *deterministic* critical pair.

It is straightforward to see that because of the symmetric probabilistic assignments of stimuli to categories, the optimal boundary for partitioning the space into response regions was the same linear boundary illustrated in Figure 4. Because decision-boundary theory assumes that mean RT is based solely on distance from this boundary, it therefore predicts equal mean RTs for the probabilistic and deterministic critical stimuli. Intuitively, according to this theory, the observer has established a simple (nonverbal) rule for classifying objects, formalized in terms of the placement of the

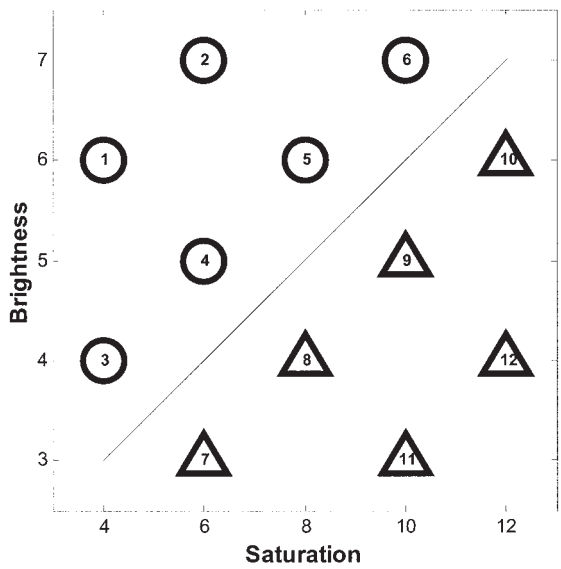


Figure 4. Schematic illustration of the design in Experiments 1 and 2. Circles are members of Category A, and triangles are members of Category B. The solid diagonal line is the optimal decision boundary for dividing the space into category-response regions. Across conditions, either Stimulus Pair 4/8 or Stimulus Pair 5/9 was assigned probabilistically to the categories.

boundary. Memories for probabilistic assignments of exemplars do not influence the application of the rule.

By contrast, the exemplar-RW model predicts that the probabilistic critical stimuli will be classified with lower accuracy and with slower response speed than the deterministic critical stimuli. For example, in Condition 4/8, in cases in which Stimulus 4 is presented and tokens of Exemplar 4 are retrieved from memory, .75 of the individual steps in the random walk will move in the direction of Category A, but .25 of the steps will move in the direction of Category B. Presentations of the deterministic critical stimuli will result in more consistent steps of the random walk, thereby leading to higher accuracy and shorter RTs.

Furthermore, the design is highly diagnostic because it relies on the collection of both accuracy and RT data. As noted previously, in much past work involving, for example, the testing of bivariate normal categories, subjects have been observed to respond in near-deterministic fashion, classifying all stimuli to one side of the decision boundary into one category and all stimuli to the other side of the boundary into the contrast category. Although the exemplar-RW model can account for this pattern of near-deterministic responding by setting the random-walk criteria at a sufficiently large magnitude, it would still predict a large effect of the probabilistic feedback assignments on the observed RT data.

It is interesting to note that this design also provides a strong contrast between the predictions of the exemplar-RW and prototype-RW models. It turns out that with the current probabilistic assignments and stimulus spacings, the centroids of each category are equidistant to the probabilistic and deterministic critical pairs. Thus, like decision-boundary theory, the prototype-RW model predicts identical choice probabilities and RTs for the probabilistic and deterministic critical pairs.

Finally, we comment on a few remaining aspects of the experimental design. We refer to Stimuli 1, 2, 11, and 12 in Figure 4 as the *far* stimuli (because they are far from the decision boundary). All three models predict that the far stimuli will be classified with the highest accuracy and fastest response speed. These stimuli were included in the design to check that the same basic distance-from-boundary effects observed in previous work would also be observed in the present experimental setting. We refer to Stimuli 3, 6, 7, and 10 in Figure 4 as the *edge* stimuli. One reason for including the edge stimuli in the design was to motivate subjects to establish the diagonal linear boundary across the range of the perceptual space. (Without the edge stimuli, subjects could learn the classification by forming a single-dimension rule and remembering a single exception.) In addition, the results for the edge stimuli provide additional constraints for quantitative model fitting. As we show, because the edge stimuli are more distant from the prototypes than are the centrally located critical stimuli, the prototype-RW model predicts much lower accuracies and longer RTs for them. By contrast, the exemplar-RW and decision-boundary models predict smaller differences in performance between the deterministic critical pairs and the edge stimuli.

We tested highly practiced subjects in the present experiment. Our aim was to test for effects of the probabilistic exemplar assignments on experienced performance rather than simply on initial learning. In addition, our goal was to conduct quantitative model fitting at the individual-subject level, so sufficient data needed to be collected for each individual subject.

Method

Subjects. The 16 subjects who participated in the speeded classification task were recruited from the Indiana University Bloomington community. Each subject received $8 per 1-hr session and participated in five sessions. A $15 bonus was promised to the 3 subjects with the best overall performance in the experiment. All subjects had normal or corrected-to-normal vision, and all claimed to have normal color vision. None of the subjects was aware of the issues under investigation in the experiment. Following the main experiment, an additional group of 39 subjects, recruited from the same population, participated in a similarity-scaling experiment.

Stimuli. The 12 color stimuli were created by scanning a set of Munsell color chips into a computer. According to the Munsell color system, the stimuli were of a constant red hue (7.5R) and varied in saturation and brightness. The saturation–brightness coordinates were as illustrated in Figure 4. Each of the colors was presented as a 2-in. (5.08-cm) square on a black background. The colors were displayed on 15-in. (38.10-cm) monitors.

Procedure. The colors were divided into two categories, as illustrated in Figure 4. Colors 4, 5, 8, and 9 were defined as the critical stimuli. In Condition 4/8, Colors 4 and 8 received probabilistic feedback—that is, they received feedback consistent with their assigned category with probability .75 and the opposite feedback with probability .25. In Condition 5/9, Colors 5 and 9 received the probabilistic feedback. All other colors received deterministic feedback.

Because the central question in this research focused on the results for the critical stimuli, to increase statistical power, we presented the individual critical stimuli with higher probability than the individual remaining stimuli. On each trial, with probability .50, 1 of the 4 critical stimuli was displayed, with its associated feedback determined randomly in accordance with the constraints described above. Likewise, on each trial, with probability .50, 1 of the 8 remaining stimuli was displayed. Note that the increased absolute frequency of the critical stimuli does nothing to change the form or placement of the optimal decision boundary. Also, because the deterministic and probabilistic critical stimuli were presented with the same absolute frequency, this factor was held constant for these stimulus pairs.

On each trial, a fixation point flashed on the center of the computer screen for 500 ms. After the fixation point disappeared, a color appeared immediately, centered on the location of the fixation. The observer made a response by pressing one of two appropriately labeled buttons on the computer keyboard (*F* for Category A and *J* for Category B). The response was followed by 1 s of feedback in which the word *CORRECT* or *INCORRECT* was displayed on the screen. The color remained on the screen for the full duration of the feedback. There was a 500-ms intertrial interval. Subjects were instructed to rest their index fingers on the appropriate response buttons throughout the testing session and to respond as quickly as possible while keeping errors to a minimum. The subjects were informed that the monetary bonus was based on a combination of short RTs and high accuracy. The instructions informed subjects that *this is a difficult task, and in some conditions it may not be possible to achieve perfect accuracy*. Other than this statement, the instructions provided no information that probabilistic feedback was assigned to some of the stimuli.

There were 850 trials per session, and each subject completed five sessions, one session per day. Thus, each subject contributed a total of 4,250 trials; across all subjects, a total of 68,000 responses were collected. The subjects were given the opportunity to take a short break after completing each fourth of an experimental session. Half of the subjects participated in Condition 4/8, and the other half participated in Condition 5/9.

In the similarity-scaling experiment, the independent group of 39 subjects provided similarity ratings for all pairs of the colors. Each subject participated in a single session consisting of 10 blocks of all 66 unique color pairs. The order of presentation of the pairs was randomized within each block. The subjects judged the similarity between each pair of colors

by using a 9-point scale ranging from 1 (*least similar*) to 9 (*most similar*). The subjects were instructed to use the full range of the scale.

Results

Similarity-scaling experiment. The main purpose of collecting the similarity judgments was to verify that the scanned colors maintained the same basic psychological structure as assumed in the Munsell scaling solution. We analyzed the mean similarity judgments by using the simple euclidean model of the ALSCAL statistical package. The resulting 2-D solution, illustrated in Figure 5, yielded a stress of .036 and accounted for 99.2% of the variance in the mean similarity ratings. Although the MDS solution derived from the similarity ratings was noisy, it displayed the same basic structure as found in the Munsell scaling. An important result revealed by the MDS solution, however, is that Pair 5/9 was somewhat more discriminable than was Pair 4/8. Indeed, the mean similarity rating for Pair 4/8 (7.83) was significantly greater than that for Pair 5/9 (7.45), $t(38) = 2.26$, $p < .05$. This stimulus-specific difference between Critical Pairs 4/8 and 5/9 needs to be considered in interpreting the results from the speeded classification experiment.

In our subsequent theoretical analyses, we used both the Munsell scaling and the present MDS solution in fitting the formal models to the speeded classification data. For all of the models, the Munsell scaling yielded better fits. Also, the relative performance of the models remained the same, regardless of the scaling solution that was used. Because none of our conclusions were changed, we report only the results that made use of the standard Munsell scaling.

Speeded classification. The first day of classification testing was considered practice, and these data were not included in the analyses. Any RT, as well as its associated response, that was shorter than 100 ms or was more than 3 standard deviations above or below the mean for that item type was omitted from further analyses. This procedure led to the omission of less than 2% of the experimental trials. The ensuing statistical analyses that we report were all conducted on the raw choice probability and mean RT data. Analyses of transformed data (i.e., an arcsine transform of the probability data and a logarithmic transform of the mean RT data) led to the same conclusions.

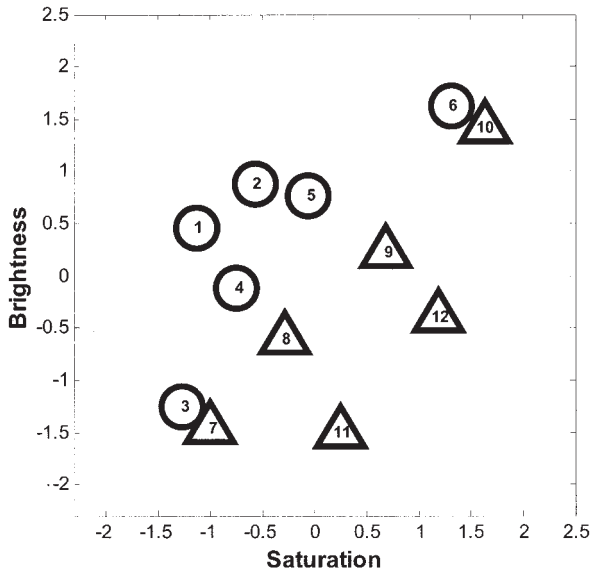


Figure 5. 2-D scaling solution for the colors derived from the similarity ratings in Experiment 1.

The choice probability data for each color for each individual subject are reported in Appendix C. The mean RT data, computed across both correct and incorrect responses, are reported in Appendix D.

The overall trends are reported in Tables 1 and 2. Note that for the probabilistic pairs, a response was defined as *correct* if the subject classified the color in accordance with the strategy of an ideal observer. For example, in Condition 4/8, regardless of the feedback provided on a given trial, a correct response was defined to occur if the subject classified Color 4 into Category A. Inspection of Table 1 reveals that subjects classified the deterministic critical pairs with higher accuracy ($M = .892$) than they did the probabilistic critical pairs ($M = .855$).

We analyzed the data for the critical pairs by using a 2 × 2 mixed-model analysis of variance (ANOVA) in which condition (4/8 vs. 5/9) was the between-subjects factor and type of feedback (probabilistic vs. deterministic) was the within-subject factor. There was a main effect of feedback, $F(1, 14) = 4.95$, $MSE = 0.002$, $p = .043$; but there was no main effect of condition, $F(1, 14) = 0.17$, $MSE = 0.005$, $p = .688$, and no Condition × Feedback interaction, $F(1, 14) = 1.46$, $MSE = 0.002$, $p = .247$. The main effect of feedback confirms our observation that the deterministic pairs were classified more accurately than were the probabilistic pairs. In addition, 13 of the 16 subjects showed more accurate responding for the deterministic pairs than for the probabilistic pairs. These results are in accord with the prior qualitative predictions from the exemplar-RW model, and they challenge the predictions from the linear decision-boundary and prototype-RW models.

As is also shown in Table 1, the far stimuli were classified much more accurately ($M = .976$) than were the deterministic critical pairs, $t(15) = 7.06$, $p < .001$. This result is as predicted by all three models. The edge stimuli were classified less accurately ($M = .854$) than were the deterministic critical pairs, $t(15) = -3.36$, $p < .01$. As we show below, although the three models correctly predict this direction of results, they differ in their predictions of the magnitude of the effect. We consider these results more fully in the *Theoretical Analysis* section.

Regarding the RTs, inspection of Table 2 reveals that subjects classified the deterministic critical pairs more quickly ($M = 602.6$ ms) than they did the probabilistic critical pairs ($M = 618.1$ ms), although the effect was not a large one. A 2 × 2 mixed-model ANOVA of the RT data revealed a marginal main effect of feedback, $F(1, 14) = 1.75$, $MSE = 1{,}091.2$, $p = .10$ (one-tailed); a marginal Condition × Feedback interaction, $F(1, 14) = 2.46$, $MSE = 1{,}091.2$, $p = .07$ (one-tailed); and no main effect of condition, $F(1, 14) = 0.11$, $MSE = 31{,}520.8$, $p = .74$. Although the effect of feedback did not reach conventional levels of statistical significance, it went in the same direction as that observed for the accuracy data (i.e., showing an advantage for the deterministic over the probabilistic critical pairs). If the accuracy and RT data sets are treated independently, then the joint probability of observing F statistics this extreme (in the predicted direction) if there were truly no effect is only .003. We believe that a reasonable conclusion is that the results comparing performance on the prob-

abilistic and deterministic critical pairs support the qualitative predictions from the exemplar-RW model. However, the marginal statistical results involving the RTs cast some doubt on the predictions from this model as well, and we pursued this issue in Experiment 2.

Note that the statistical interaction between condition and feedback reflects a stimulus-specific effect in which, overall, Pair 5/9 tended to be classified more rapidly than Pair 4/8. This result is unsurprising given the results of the similarity-scaling experiment, which indicated that Pair 5/9 was more discriminable than was Pair 4/8.

Finally, as shown in Table 2, the mean RTs for the far stimuli (531.8 ms) were clearly shorter than those for the deterministic critical pairs, $t(15) = 6.60$, $p < .001$. The edge stimuli were classified more slowly on average (631.6 ms) than were the deterministic critical pairs, $t(15) = 1.88$, $p = .04$ (one-tailed).

In a further analysis, we broke down the data according to whether they were collected during Sessions 2 and 3 or Sessions 4 and 5. Note that given the nature of the design, an ideal observer would perform best by ignoring the probabilistic feedback and assuming instead that the correct feedback is that provided for each stimulus on the majority of the trials. Thus, we hypothesized that the effects of the probabilistic feedback might grow weaker during the later sessions of testing. Analysis of the data, however, showed little effect of session (except for an overall speeding of all responses due to generalized practice effects). If anything, the effects of the probabilistic feedback grew stronger during the later sessions, although the results did not approach statistical significance.

Theoretical Analysis

As a source of converging evidence, we conducted tests of the models' ability to quantitatively fit the individual-subject choice probability and mean RT data. Although the qualitative contrasts described in the previous section favored the predictions from the exemplar-RW model, our view is that quantitative comparisons are also of fundamental importance. For example, suppose that the exemplar-RW model were to predict a quantitative performance advantage for the deterministic pairs relative to the probabilistic pairs that was far greater in magnitude than the observed advantage. Such a result would be reflected in a poor overall quantitative fit, thereby casting doubt on the modeling ideas. Likewise, the quantitative tests consider the ability of the models to capture the complete constellation of results in the data, not solely the single qualitative contrast that was the focus of the design. Thus, if the exemplar-RW model captures only the single qualitative contrast involving the deterministic and probabilistic pairs but fails badly to fit other aspects of the data, this would cast doubt on the model as well.

Table 1
Proportions of Correct Classifications for the Main Stimulus Types in Each Condition of Experiment 1

	Condition		
Stimulus type	4/8	5/9	Average
Prob	.850	.859	.855
Det	.907	.876	.892
Edge	.854	.854	.854
Far	.983	.969	.976

Note. Prob = probabilistic critical pair; Det = deterministic critical pair; Edge = edge stimuli; Far = far stimuli.

Table 2
Mean Response Times (in Milliseconds) for the Main Stimulus Types in Each Condition of Experiment 1

	Condition		
Stimulus type	4/8	5/9	Average
Prob	616.6	619.5	618.1
Det	582.8	622.4	602.6
Edge	620.8	642.4	631.6
Far	525.5	538.1	531.8

Note. Prob = probabilistic critical pair; Det = deterministic critical pair; Edge = edge stimuli; Far = far stimuli.

Recall that each model had six free parameters.[3] Each individual subject's data set had 24 freely varying data entries, 12 choice probabilities, and 12 mean RTs. We fitted the three models to the individual-subject choice probability and mean RT data by searching for the values of the free parameters that minimized a weighted sum-of-squared-deviations (WSSD) statistic. Each squared deviation (between predicted and observed data values) was weighted by the inverse of the squared standard error of that data value.[4] Thus, highly variable data values contribute less to the WSSD than do less variable data values. An important advantage of using the WSSD statistic is that it basically places the choice probability and RT data on the same scale, with both contributing roughly equally to the overall goodness-of-fit evaluation. Although an improved fit statistic might involve the use of a maximum-likelihood criterion, we found the derivation of a joint likelihood statistic for the choice probability and RT data to be intractable. Finally, to guard against local minima, we used multiple starting configurations in the parameter searches. The predicted choice probabilities and mean RTs for each color and each individual subject are reported along with the observed data in Appendixes C and D.

The WSSD results from the three models are reported for each individual subject in Table 3. We compared the fits of the exemplar-RW and decision-boundary models by using a 2×2 ANOVA with condition (4/8 vs. 5/9) as a between-subjects factor and model as a within-subject factor. Although the mean fit value for the exemplar-RW model (174.9) was better than that for the

[3] For all three models, the lower limit of the mean residual-time parameter μ was set at 100 ms. Following previous work, for the exemplar-RW and prototype-RW models, the decision criteria $+A$ and $-B$ were allowed to be real-valued in application of the analytic prediction equations. With regard to predicting choice probabilities and mean RTs, this procedure provides a close approximation to assuming that there is a probabilistic mixture of integer-valued decision-criterion settings across trials.

[4] To implement the WSSD statistic, observed choice probabilities equal to 0 were set equal to $\frac{1}{2}N$ instead, where N is the number of observations on which the choice probability is based. Likewise, observed choice probabilities equal to 1 were set equal to $(2N - 1)/2N$ instead. Otherwise, the inverse of the squared standard error would be equal to infinity and the WSSD statistic undefined.

decision-boundary model (212.1), this difference was not statistically significant, $F(1, 14) = 2.27$, $MSE = 4{,}882.1$, $p = .154$. The Condition × Model interaction also failed to reach statistical significance, $F(1, 14) = 2.18$, $MSE = 10{,}658.0$, $p = .162$. The trend, however, was that the exemplar-RW model had an advantage in fitting the Condition 4/8 data, whereas the decision-boundary model fitted the Condition 5/9 data somewhat better. As we show below, the main reason for this pattern is that the exemplar-RW model predicts a performance advantage for the deterministic critical pairs over the probabilistic critical pairs. This predicted advantage was observed in Condition 4/8, but there was little difference between the two types of pairs in Condition 5/9. The reason, as noted above, is that beyond the effect of the probabilistic feedback assignments, there was also a stimulus-specific effect in which Pair 5/9 was processed more efficiently overall than was Pair 4/8. Again, this result seems reasonable given the results of our similarity-scaling experiment.[5]

Finally, as is also shown in Table 3, the exemplar-RW model provided a far better fit to the individual-subject data ($M = 305.9$) than did the prototype-RW model. A 2 × 2 ANOVA with condition and model as factors revealed a significant main effect of model, $F(1, 14) = 17.04$, $MSE = 8{,}061.6$, $p < .001$. The Condition × Model interaction did not approach statistical significance. The advantage in fit for the exemplar-RW model was observed for 14 of the 16 subjects.

To provide some sense of the reason for these fit differences, in Table 4 we report collapsed predictions from the models for the four main types of stimuli. (Although the models were fitted to the individual-subject data, the aggregated predictions in the table were obtained by averaging across the results from the individual subjects.)

As can be seen in Table 4, the linear decision-boundary model predicts nearly identical choice probabilities and mean RTs for the probabilistic and deterministic critical pairs. By contrast, the observed data show an overall advantage for the deterministic pairs. Recall that if subjects adopted the optimal boundary illustrated in Figure 4, the predicted choice probabilities and mean RTs would be identical. By allowing the slope and *y*-intercept of the linear boundary to be free parameters, the model can deviate slightly from this strong prediction, but the amount of adjustment is insufficient to account for the observed data. The prototype-RW model has the same limitation as does the linear decision-boundary model with respect to the critical pairs. It predicts essentially identical choice probabilities and mean RTs for the probabilistic and deterministic critical pairs, thereby failing to account for the observed differences in the data. In addition, the prototype-RW model predicts a performance advantage for the deterministic critical pairs over the edge stimuli that is much too large, especially in the choice probability data. The reason that the prototype-RW model predicts this advantage is that the critical-pair stimuli lie much closer to their category prototypes than do the edge stimuli.

The exemplar-RW model comes closer overall to predicting the main trends in the data than do the alternative models. First, it predicts well the magnitude of the accuracy advantage for the deterministic pairs over the probabilistic pairs. Second, it makes the correct qualitative prediction of an RT advantage for the deterministic pairs over the probabilistic pairs, although it overes-

Table 3
Summary Weighted Sum-of-Squared-Deviations Fits of Each Model to the Individual-Subject Data From Experiment 1

	Model		
Subject	Exemplar-RW	Decision bound	Prototype-RW
1	34.8	34.9	180.8
2	277.4	363.0	250.9
3	92.0	95.4	157.9
4	122.0	285.1	570.7
5	78.6	87.9	291.2
6	490.9	839.9	551.1
7	202.3	175.0	226.4
8	128.4	134.8	233.1
9	264.9	324.2	370.5
10	171.5	216.7	147.8
11	166.5	160.1	411.1
12	150.3	239.3	186.4
13	143.5	79.5	406.9
14	112.1	84.3	271.3
15	95.0	73.9	174.1
16	268.2	199.6	464.5
M	174.9	212.1	305.9

Note. Subjects 1–8 participated in Condition 4/8; Subjects 9–16 participated in Condition 5/9. Exemplar-RW = exemplar-based random-walk model; Decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model.

Table 4
Collapsed Predictions From the Formal Models of the Main Trends in the Mean Accuracy and Response Time Data in Experiment 1

		Model		
Stimulus type	Obs.	Exemplar-RW	Decision bound	Prototype-RW
		Mean proportions correct		
Prob	.855	.868	.884	.931
Det	.892	.913	.881	.928
Edge	.854	.865	.877	.808
Far	.976	.984	.998	.984
		Mean response times (ms)		
Prob	618.1	618.2	606.7	584.2
Det	602.6	592.1	607.7	584.0
Edge	631.6	614.8	609.6	633.7
Far	531.8	528.7	527.9	534.6

Note. Obs. = observed data; Exemplar-RW = exemplar-based random-walk model; decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model; Prob = probabilistic critical stimuli; Det = deterministic critical stimuli; Edge = edge stimuli; Far = far stimuli.

[5] We had hoped that when used in combination with the derived MDS solution for the colors, the exemplar-RW model could capture this joint influence of the probabilistic feedback and differing stimulus-pair discriminabilities. However, as noted earlier, all models yielded better fits to the data when used in combination with the Munsell scaling rather than with the MDS solution derived from the similarity ratings. Our interpretation is that the overall MDS solution is too noisy, relative to the Munsell scaling solution, to yield improved quantitative fits to the complete sets of choice probability and mean RT data.

timates the magnitude of the observed difference. (We pursue the latter issue in Experiment 2.) Third, the exemplar-RW model predicts well the magnitude of the performance advantage for the deterministic critical pairs over the edge stimuli. The reason for this prediction is that the deterministic critical pairs are centrally located in the similarity space, and they are presented with higher absolute frequency than the edge stimuli. Thus, the deterministic critical pairs give rise to somewhat higher summed-activation values than do the edge stimuli, and this accounts for their predicted performance advantage. The exemplar-RW model also yields good quantitative predictions of the accuracy and mean RT associated with the far stimuli.

Discussion

The first main result of importance is that subjects classified the deterministic critical pairs with higher accuracy than they did the probabilistic critical pairs. This result is in accord with the predictions from the exemplar-RW model, and it challenges the predictions from the linear decision-boundary and prototype-RW models. The mean RTs for the deterministic critical pairs were also shorter than those for the probabilistic critical pairs, although this result did not reach conventional levels of statistical significance. Taken together, however, the choice probability and RT data favor the predictions from the exemplar-retrieval model. Likewise, the quantitative model-fitting results support the predictions of the exemplar-RW model over those of the linear decision-boundary and prototype-RW models. However, the fit differences between the linear decision-boundary and exemplar-RW models again did not reach conventional levels of statistical significance. The main reason seems to be that the magnitude of the RT difference between the probabilistic and deterministic pairs was not as large as that predicted by the exemplar-RW model. Because the initial motivation of our research was to investigate an effect of the probabilistic exemplar assignments on classification RTs, and because the RT results from Experiment 1 were marginal, we decided to pursue this issue further in Experiment 2.

Experiment 2

The main purpose of Experiment 2 was to test further whether probabilistic assignments of exemplars to categories might indeed affect the time course of classification decision making. The key idea in the experiment was to induce subjects to place greater emphasis on accuracy than they had in Experiment 1 while maintaining the general context of a speeded classification situation. There are a couple of ways in which an increased emphasis on accuracy might affect the random-walk decision process. First, it might lead subjects to use stricter decision criteria (i.e., to increase the magnitude of the criteria $+A$ and $-B$ in the random walk). According to the exemplar-RW model, if the decision criteria are moved outward, it should take a greater number of steps, on average, to complete the random walk. Thus, according to theory, any true differences in classification RTs among the stimulus types would be magnified relative to what was observed in Experiment 1. A second possibility is that subjects would work harder to extract more fine-grained perceptual information from the stimulus displays (which would be reflected in an increase of the value of the overall sensitivity parameter c). Presumably, this increased processing effort would be reflected in an increase in the time required to take each individual step in the random walk. Again, according to theory, any true differences in classification RTs between the different stimulus types would thereby be magnified.

A possible drawback of inducing longer RTs is that there might be more noise in the observed RT data. In addition, the accuracy data might approach a ceiling, thereby removing an important source of information for distinguishing among the models. Nevertheless, because the motivating theme of our initial investigation had focused on classification RTs, the idea seemed like a reasonable one to pursue.

In Experiment 1, our instructions placed emphasis on both speed and accuracy. In Experiment 2, to give greater emphasis to accuracy and to possibly magnify RT differences, we paid subjects monetary bonuses for making correct responses. However, to maintain the general context of a speeded classification situation, each trial had an RT deadline of 5 s. Failure to meet the deadline was counted as an incorrect response in calculating the bonus. Our intent was to choose a deadline sufficiently long that no real time pressure was exerted, yet the general context of a speeded classification situation was maintained. In all respects except for the instructions, Experiment 2 was the same as Experiment 1.

Method

Subjects. There were 10 new subjects recruited from the Indiana University Bloomington community. Half participated in Condition 4/8, and the other half participated in Condition 5/9. Each subject received $8 per 1-hr session, plus monetary bonuses for good performance (described below). All subjects had normal or corrected-to-normal vision, and all claimed to have normal color vision. None of the subjects was aware of the issues under investigation in the experiment.

Stimuli. The stimuli were the same as in Experiment 1.

Procedure. All aspects of the procedure were the same as in Experiment 1 except for the instructions regarding the monetary bonuses. Subjects were informed that each time they made a correct response, $0.01 would be added to their monetary bonus, whereas incorrect responses would lead to $0.01 reductions. In addition, failure to meet a 5-s RT deadline would also result in a $0.01 reduction in the bonus, regardless of whether the response was correct. The accumulated bonus on each trial was displayed on the bottom of the computer screen during the period in which feedback was provided.

Results

As was the case in Experiment 1, the first day of classification testing was considered practice, and these data were not included in the analyses. Also, any RT, as well as its associated response, that was shorter than 100 ms or was more than 3 standard deviations above or below the mean for that item type was omitted from further analyses. This procedure led to the omission of less than 2% of the experimental trials.

As a manipulation check on the instructions, we examined the mean accuracy and RT data and compared them with the results observed in Experiment 1. Averaged across all stimuli, and using subjects as the unit of analysis, mean accuracy was significantly higher in Experiment 2 (.951) than in Experiment 1 (.901), $t(24) = 3.02$, $p < .01$. In addition, mean RTs were significantly longer in Experiment 2 (712.8 ms) than in Experiment 1 (591.2 ms), $t(24) = 2.36$, $p < .05$. Thus, our modified instructions had the desired general effect of increasing accuracy and slowing down overall response speed.

The individual-subject data from Experiment 2 are reported in Appendixes E and F. The overall trends are reported in Tables 5 and 6 for the accuracy and mean RT data, respectively. The patterns of the data are identical to those observed in Experiment 1. The main difference is that the magnitude of the RT differences between the critical stimulus types is greater than it was in Experiment 1.

As can be seen in Table 5, mean accuracy was once again greater for the deterministic critical pairs (.948) than it was for the probabilistic critical pairs (.891). There was also a stimulus-specific effect in which overall accuracy for Pair 5/9 ($M = .945$) was greater than overall accuracy for Pair 4/8 ($M = .894$). We conducted a 2 × 2 ANOVA using condition (4/8 vs. 5/9) and type of feedback (probabilistic vs. deterministic) as factors. The analysis revealed a main effect of feedback, $F(1, 8) = 19.10$, $MSE = 0.001$, $p = .002$, reflecting the superiority of the deterministic pairs over the probabilistic pairs, and a significant Condition × Feedback interaction, $F(1, 8) = 15.30$, $MSE = 0.001$, $p = .004$, reflecting the stimulus-specific advantage of Pair 5/9 over Pair 4/8. There was no main effect of condition, $F(1, 8) = 1.14$, $MSE = 0.003$, $p = .317$.

We also conducted a 2 × 2 ANOVA of the mean RT data. The most important result, shown in Table 6, was that mean RT was significantly shorter for the deterministic critical pairs (731.2 ms) than it was for the probabilistic critical pairs (799.7 ms), $F(1, 8) = 7.49$, $MSE = 3{,}136.5$, $p = .026$. The analysis also revealed a significant Condition × Feedback interaction, $F(1, 8) = 8.71$, $MSE = 3{,}136.5$, $p = .018$, reflecting the stimulus-specific RT advantage of Pair 5/9 over Pair 4/8. There was no main effect of condition, $F(1, 8) = 0.60$, $MSE = 40{,}342.7$, $p = .46$. The effects of the probabilistic exemplar assignments on both the accuracy and RT data are in strong agreement with the qualitative predictions from the exemplar-RW model.

As was the case in Experiment 1, mean accuracy for the far stimuli (.995) was significantly greater than that for the deterministic critical pairs, $t(9) = 6.27$, $p < .001$; and mean RT for the far stimuli (624.6 ms) was significantly shorter than that for the deterministic critical pairs, $t(9) = 8.72$, $p < .001$. Thus, a strong distance-from-boundary effect was again clearly present in the data. Although the edge stimuli had slightly lower accuracies ($M = .942$) and longer RTs ($M = 748.4$ ms) than did the deterministic critical pairs, these differences were not statistically significant for the accuracy, $t(9) = 0.68$, $p > .10$, or the RT data, $t(9) = 0.83$, $p > .10$.

Further analysis revealed that the same patterns of accuracy and RT data held across both Sessions 2 and 3 and Sessions 4 and 5 of testing. Thus, although an ideal observer would respond more accurately and more rapidly by ignoring the probabilistic feedback, the manipulation continued to exert an influence, even after 5 days of testing.

Table 5
Proportions of Correct Classifications for the Main Stimulus Types in Each Condition of Experiment 2

	Condition		
Stimulus type	4/8	5/9	Average
Prob	.852	.929	.891
Det	.961	.935	.948
Edge	.935	.949	.942
Far	.995	.995	.995

Note. Prob = probabilistic critical pair; Det = deterministic critical pair; Edge = edge stimuli; Far = far stimuli.

Table 6
Mean Response Times (in Milliseconds) for the Main Stimulus Types in Each Condition of Experiment 2

	Condition		
Stimulus type	4/8	5/9	Average
Prob	801.9	797.6	799.7
Det	659.4	803.0	731.2
Edge	716.3	780.5	748.4
Far	583.6	665.7	624.6

Note. Prob = probabilistic critical pair; Det = deterministic critical pair; Edge = edge stimuli; Far = far stimuli.

Theoretical Analysis

Fits of models. We fitted the models to the individual-subject classification data in Experiment 2 by using the same procedure as in Experiment 1. The individual-subject predictions are reported along with the observed data in Appendixes E and F.

In Table 7, we report the individual-subject fit values achieved by each of the models. The exemplar-RW model again provided the best overall fit to the individual-subject data. Furthermore, a 2 × 2 ANOVA revealed that the mean WSSD yielded by the exemplar-RW model (205.7) was significantly smaller than the one yielded by the decision-boundary model (270.7), $F(1, 8) = 8.25$, $MSE = 2{,}558.4$, $p = .021$. The Condition × Model interaction was also significant, $F(1, 8) = 14.12$, $MSE = 2{,}558.4$, $p = .006$. The interaction reflects the stimulus-specific effect involving Pairs 4/8 and 5/9: The fit of the exemplar-RW model was substantially better than that of the decision-boundary model in Condition 4/8, whereas it fared slightly worse in Condition 5/9. Overall, the exemplar-RW model yielded a better fit than did the decision-boundary model for 7 of the 10 subjects.

The exemplar-RW model provided substantially better fits to the individual-subject data than did the prototype-RW model ($M = 520.1$), $F(1, 8) = 70.17$, $MSE = 7{,}042.4$, $p < .001$. In this case, the Condition × Model interaction did not approach statistical significance. Indeed, the exemplar-RW model outperformed the prototype-RW model for all 10 subjects.

To provide some sense of the reason for these model-fit results, in Table 8 we report the collapsed predictions from the models for the four main stimulus types. The patterns of predictions are the same as those seen in Experiment 1. Both the linear decision-boundary model and the prototype-RW model predict virtually identical accuracies and mean RTs for the probabilistic and deterministic critical pairs. By contrast, the exemplar-RW model correctly predicts the performance advantage, in both accuracy and mean RT, observed for the deterministic pairs. It does a good job of predicting performance for the edge and far stimuli as well.

Distributional analyses of extended decision-boundary hypotheses. In this section, we consider various extended versions of the decision-boundary model that might allow this approach to account for the effect of the probabilistic exemplar assignments.

One possibility is that the probabilistic exemplar assignments might give rise to uncertainty effects in subjects' perceptions of the stimuli.[6] In past work, Ashby and Maddox (1994) proposed to model uncertainty effects in terms of increased variances of the perceptual distributions associated with each stimulus. Suppose that the perceptual distributions associated with the probabilistic pairs had greater variances than those associated with the deterministic pairs. One consequence is that there would be reduced accuracy for the probabilistic pairs, because a greater proportion of their perceptual distributions would overflow into the incorrect category-response region. A second consequence is that mean RT would be lengthened for the probabilistic pairs. The reason is that compared with the deterministic pairs, a greater proportion of the percepts associated with the probabilistic pairs would lie close to the decision boundary. It is critical to note that the increase in variance would also result in an increased proportion of the percepts being located farther from the boundary. However, because RT is an exponentially decreasing function of distance from the boundary, mean RT would still tend to be longer when averaged across all percepts.

This increased-variance hypothesis makes another strong prediction, however. Specifically, because some of the percepts are even farther from the boundary, the very shortest RTs associated with the probabilistic pairs should be shorter than the very shortest RTs associated with the deterministic pairs (for a similar argument in a related context, see Nosofsky & Palmeri, 1997a, pp. 1032–1033). Thus, the uncertainty hypothesis can be tested by conducting analyses on the fine-grained RT distribution data. We considered all subjects who did indeed display longer mean RTs for the probabilistic pairs than for the deterministic pairs. (We focused on only these subjects because the goal was to test the perceptual-variance explanation of slowed responding on the probabilistic pairs.) We then extracted only the shortest 5% of the RTs from the complete RT distributions associated with these stimuli. In Experiment 1, among those subjects who were slower overall on the probabilistic pairs, the mean of the shortest 5% of RTs was 378.6 ms for the probabilistic pairs and 368.5 ms for the deterministic pairs. In Experiment 2, the means were 455.5 ms and 450.7 ms, respectively. This pattern is in the opposite direction of what is predicted by the uncertainty hypothesis. Furthermore, across both experiments, there were only 2 subjects for whom the shortest RTs associated with the probabilistic pairs were shorter than those associated with the deterministic pairs, and the differences here were small. Therefore, the distributional analysis provides little support for the idea that the effects of the probabilistic assignments can be explained in terms of increased perceptual variance.

Another possibility that we considered is that the location of the decision boundary might be altered systematically because of the probabilistic exemplar assignments. For example, because of the inconsistent feedback, perhaps a decision boundary would be developed with a suboptimal slope such that it was located midway between the deterministic stimuli but very close to one of the probabilistic stimuli (and so, very far from the other probabilistic stimulus). Once again, however, such a model predicts that the very shortest RTs would be associated with the probabilistic stimuli, not the deterministic ones, and our distributional analysis provides no evidence in favor of this prediction. In addition, our quantitative model-fitting analysis allowed the slope and the *y*-intercept of the linear boundary to be free parameters, but the model tended to fare worse than did the exemplar-RW model in fitting the individual-subject data.

Table 7
Weighted Sum-of-Squared-Deviations Fits of Each Model to the Individual-Subject Data From Experiment 2

	Model		
Subject	Exemplar-RW	Decision bound	Prototype-RW
1	154.6	381.8	676.0
2	165.3	322.3	554.9
3	91.2	140.0	462.6
4	133.9	219.2	349.1
5	167.1	398.7	450.0
6	81.9	92.6	188.4
7	327.0	270.4	760.8
8	401.0	468.2	619.6
9	341.2	262.6	626.9
10	194.0	151.2	512.7
M	205.7	270.7	520.1

Note. Subjects 1–5 participated in Condition 4/8; Subjects 6–10 participated in Condition 5/9. Exemplar-RW = exemplar-based random-walk model; Decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model.

Table 8
Collapsed Predictions From the Formal Models of the Main Trends in the Mean Accuracy and Response Time Data in Experiment 2

		Model		
Stimulus type	Obs.	Exemplar-RW	Decision bound	Prototype-RW
		Mean proportions correct		
Prob	.891	.926	.952	.986
Det	.948	.975	.956	.986
Edge	.942	.958	.952	.928
Far	.994	.996	1.000	.998
		Mean response times (ms)		
Prob	799.7	786.1	740.4	693.8
Det	731.2	711.1	740.0	693.6
Edge	748.4	742.4	740.9	771.9
Far	624.7	623.3	620.6	640.7

Note. Obs. = observed data; Exemplar-RW = exemplar-based random-walk model; decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model; Prob = probabilistic critical stimuli; Det = deterministic critical stimuli; Edge = edge stimuli; Far = far stimuli.

[6] In past work (Ashby & Maddox, 1994), uncertainty effects in the perceptual representation were theorized to occur because of the probability with which individual stimuli were presented. It is unclear whether such uncertainty effects in perception would also be expected to occur on the basis of response feedback that is received. Nevertheless, we make allowance for this possibility in considering the predictions from decision-boundary theory.

General Discussion

Three of the major approaches to modeling the nature of category representation and decision processes in perceptual classification are exemplar-retrieval, prototype, and decision-boundary models. In our view, there is already much evidence that severely challenges the predictions from prototype models (e.g., Ashby & Maddox, 1992; Medin & Schaffer, 1978; Nosofsky, 1987; Nosofsky & Zaki, 2002), and the present results add to that body of evidence. By contrast, despite their vast conceptual differences, exemplar and decision-boundary models make surprisingly similar predictions in a wide variety of paradigms. We focus this General Discussion, therefore, on a review of recent comparisons between decision-boundary and exemplar-retrieval models. In particular, in our view, evidence is beginning to mount that challenges strong versions of decision-boundary theory as well. The results from the current experiments provide important converging evidence along these lines.

Exemplar-Retrieval and Decision-Boundary Models

One source of evidence involves classification performance in situations that test complex category structures. Strong versions of decision-boundary theory posit that people use either linear or quadratic boundaries for partitioning a perceptual space into response regions. The main motivation for this hypothesis is the assumption that numerous categories in the natural world are multivariate normally distributed. It is well known that the optimal boundary for partitioning two multivariate normal categories is always linear or quadratic in form (see Ashby & Gott, 1988). It is linear when the two categories have the same variance–covariance structure (i.e., when the two category distributions have the same size and shape); otherwise, the optimal boundary is quadratic. Because a central assumption in early versions of the theory was that people will adopt decision boundaries with an optimal form, and that the category-learning system assumes normal distributions, most early work focused on tests of linear and quadratic decision-boundary models. Indeed, such models fared very well in situations in which subjects in fact learned to classify members of bivariate normal category distributions (Ashby & Maddox, 1992; Maddox & Ashby, 1993).

Two studies, however, provided important challenges to this strong version of decision-boundary theory. First, McKinley and Nosofsky (1995) tested subjects in designs in which the category structures were based on mixtures of normal distributions. In these designs, the optimal boundary for partitioning the space into response regions was highly nonquadratic; instead, it was more complex in form (see McKinley & Nosofsky, 1995, Figure 3). Furthermore, according to exemplar models, subjects should learn to classify stimuli in rough accordance with the use of these complex boundaries. McKinley and Nosofsky's experiments provided clear evidence that the vast majority of subjects performed in the manner predicted by the exemplar model and not according to the predictions of the linear or quadratic decision-boundary models.

Likewise, Ashby and Waldron (1999) tested subjects in designs involving category structures that used transformed normal distributions. In one experiment, if subjects assumed normal distributions, then the adopted decision boundary would be linear in form; however, the optimal boundary for partitioning the transformed categories was quadratic. In a second experiment, if subjects assumed normal distributions, the adopted decision boundary would be quadratic in form; by contrast, the optimal boundary for partitioning the transformed categories was linear. Ashby and Waldron obtained overwhelming evidence that subjects behaved as if they were using the optimal boundary for the transformed distributions (i.e., they were not assuming normal distributions when classifying the objects).

The results from McKinley and Nosofsky (1995) and Ashby and Waldron (1999) thereby provided strong challenges to versions of decision-boundary theory based on the assumption of normal distributions. They did not, however, rule out the more general idea that complex decision boundaries could be learned "online" and be approximately optimal for each individual category structure tested. Therefore, it was important to seek alternative sources of evidence for contrasting exemplar-retrieval and decision-boundary models.

As noted in the introduction, one approach was to contrast the models' predictions in speeded classification situations in which the absolute frequency of individual exemplars was manipulated (Nosofsky & Palmeri, 1997b, Experiment 2; Verguts et al., 2003). A key feature of these designs was that the distance of the objects from the optimal decision boundary was not affected by these absolute-frequency manipulations. The general result from these studies was that exemplars presented with high frequency were classified more rapidly than were low-frequency exemplars, in accordance with the predictions from exemplar-retrieval models. Again, such results pose a challenge to versions of decision-boundary theory that assume that RT is based solely on the distance of an object from the category boundary. Nevertheless, it is possible to attribute such familiarity effects to psychological factors not associated with classification decision making (e.g., encoding or surprise effects).

Another recent study that has provided a challenge to decision-boundary theory is the set of experiments reported by Rouder and Ratcliff (2004). As described above, in various designs, Rouder and Ratcliff tested subjects on unidimensional category structures in which individual exemplars were assigned probabilistically to the alternative categories. In general, the designs pitted distance-from-boundary and stimulus probability against one another such that stimuli farther from the boundary sometimes received target-category feedback with low probability. As a result, the designs placed severe constraints on the predictions from the competing models: The decision-boundary models predicted increasing categorization probabilities as a function of distance-from-the-boundary, whereas the exemplar model tended to predict that categorization responses would track the feedback probabilities. The general pattern of results from Rouder and Ratcliff's experiments was that in situations in which highly confusable stimuli were used, performance was more in accord with the predictions from decision-boundary theory than those from exemplar-retrieval theory. However, in situations involving more discriminable stimuli, the results were more in accord with the predictions from the exemplar model.

The key manipulation in our present experiments was similar in theme to the one used by Rouder and Ratcliff (2004), but it differed in important respects as well. The idea in our experiments was to manipulate stimulus probability as an experimental factor while holding the variable of distance-from-boundary constant. As a result, we achieved a fundamental qualitative contrast between

the predictions from the models that was not present in Rouder and Ratcliff's designs. The results of our experiments converge strongly with the evidence from Rouder and Ratcliff by demonstrating that at least in situations involving fairly discriminable stimuli, exemplar-retrieval models provide a better account of perceptual classification performance than do decision-boundary models. In particular, probabilistic feedback assignments exert a strong influence on classification performance that is in accord with the predictions from exemplar-retrieval models but that is not predicted by decision-boundary models. Furthermore, our results generalize Rouder and Ratcliff's findings by showing that probabilistic feedback exerts an influence in multidimensional classification domains instead of only in unidimensional ones. Finally, our results show a systematic effect of the probabilistic feedback assignments on the time course of classification decision making, not solely on choice probabilities. This approach is important because RTs often provide a window into psychological processing that is not available from consideration of choice probabilities alone.

In our view, this systematic evidence that probabilistic exemplar assignments exert a powerful influence on classification behavior is highly intriguing. In particular, the evidence points to a stubborn form of suboptimality in human performance. Given the nature of our design, subjects would have performed optimally by simply ignoring the probabilistic feedback and classifying each object into the category that received its given feedback on the majority of trials. Indeed, such a strategy could have been implemented by forming an exceedingly simple linear boundary through the perceptual space and classifying objects in accordance with the use of this boundary. Nevertheless, despite being provided with monetary payoffs for correct responses, and even after 5 days of experience with the task, subjects' behavior departed systematically from such an optimal strategy in a manner that was well predicted by the exemplar-retrieval model.

Limitations and Future Research Directions

An important limitation of the present experiments (and those of Rouder & Ratcliff, 2004) is that in all cases, a relatively small number of exemplars defined each of the categories. The main reason for the use of a small number of exemplars was that these studies had the ambitious goal of quantitatively modeling performance at the individual-stimulus level, and adequate sample sizes are needed to achieve such tests. It is important to point out that decision-boundary theorists have often conducted and/or modeled classification experiments with the same or even fewer numbers of category exemplars (e.g., Ashby & Maddox, 1994; Maddox & Ashby, 1993, pp. 60–67; Maddox & Ashby, 1996; Maddox, Ashby, & Gottlob, 1998; Thomas, 1996) and have interpreted the results within the framework of decision-boundary theory. Thus, our experiments do not seem to go outside the scope for which this theory was intended. Nevertheless, a reasonable concern is that exemplar-retrieval processes may dominate in classification only in situations in which category sizes are small. Therefore, a critical next step in this research will be to examine the role of probabilistic exemplar assignments on speeded classification in situations involving large-size category distributions.

Another aim of future research should be to test the models' RT predictions at a more fine-grained level. In the present work, we chose to analyze and model the overall mean RTs for the individual stimuli as a function of experimental condition. In the present paradigm, the models made strongly contrasting predictions of the overall mean RTs, so this level of analysis was a reasonable one to pursue. Nevertheless, more rigorous tests would also involve consideration of the complete distribution of RTs as well as any differences between correct and error RTs. In past work, the exemplar-RW model has been shown to capture well the overall form of speeded classification RT distributions (e.g., Nosofsky, 1997; Nosofsky & Palmeri, 1997a). It would be interesting, however, to test for effects of the probabilistic exemplar assignments on this more detailed aspect of performance.

Still another avenue to pursue would involve a detailed consideration of sequential effects in speeded classification data. Given the natural assumption that more recently presented exemplars have greater memory strengths than do exemplars presented in the distant past, the exemplar model predicts interesting sequence effects of probabilistic feedback. For example, the model predicts that if a Category-A stimulus received Category-B feedback on its most recent past presentation, on the current trial there should be reduced accuracy and slowed responding for that stimulus. Indeed, an analysis of the sequence effects in our data revealed precisely this predicted pattern of results. For each individual subject, we partitioned the data for the probabilistic stimuli according to whether they received correct or incorrect feedback on their most recently presented trial. In Experiment 1, for 15 of the 16 subjects, responding was more accurate and faster when the probabilistic stimuli had received correct rather than incorrect feedback on their most recent presentation. In Experiment 2, all 10 of the subjects displayed this pattern of results. It remains an open question whether sequence-sensitive versions of the exemplar model can account in quantitative detail for these effects. Such results pose an interesting challenge to the decision-boundary and prototype models as well.

Finally, as noted in the introduction, some of the major recent theories in the field of perceptual classification posit that multiple cognitive systems underlie the representation of categories. For example, according to Ashby et al.'s (1998) COVIS (competition between verbal and implicit systems) model, there are separate explicit (verbal, rule-based) and implicit (procedural) systems. The implicit system dominates in situations in which no salient verbal rule is available for classification, such as the present experiments. As suggested by Ashby and Waldron (1999), a striatal pattern classifier might underlie the implicit procedural-learning system. The general idea in this neuropsychological model is that individual stimuli are represented in a perceptual space in high-level visual areas. However, a low-resolution map of this perceptual space is then represented among striatal units. The striatal units learn to associate category labels with different regions of perceptual space.

In its current form, the striatal-classifier model of Ashby and Waldron (1999) cannot be tested because a specific learning algorithm has not been proposed. Nevertheless, the data reported in this article should provide useful constraints for the further development of this model. For example, the category structure tested in this study could be learned by using a single striatal unit (or prototype) for each category. Such a single-unit model is formally equivalent to a linear decision-boundary model (Ashby & Waldron, 1999, pp. 374–375). As already seen, however, such a model would fail to account for the effects of the probabilistic exemplar assignments observed in the present research. Accordingly, the

resolution of the striatal map would need to be more finely grained. It remains to be seen if some intermediate-resolution map—more finely grained than a single-category prototype but coarser than individual-exemplar representations—would provide important benefits in accounting for the present speeded classification data.

References

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review, 8,* 629–647.

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., & Casale, M. B. (2003). The cognitive neuroscience of implicit category learning. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 109–142). Philadelphia: John Benjamins.

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

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400.

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., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

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

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97,* 523–547.

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 538–564.

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.

Garner, W. R. (1974). *The processing of information and structure.* New York: Wiley.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

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, and Cognition, 24,* 695–711.

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

Link, S. W. (1992). *The wave theory of difference and similarity.* Hillsdale, NJ: Erlbaum.

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

Logan, G. D. (1997). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review, 103,* 603–649.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

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

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

Maddox, W. T., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics, 60,* 620–637.

Marley, A. A. J. (1992). Developing and characterizing multidimensional Thurstone and Luce models for identification and preference. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 299–333). Hillsdale, NJ: Erlbaum.

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology, 35,* 1–20.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 5,* 207–238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

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,* 87–109.

Nosofsky, R. M. (1988a). On exemplar-based exemplar representations: Reply to Ennis (1988). *Journal of Experimental Psychology: General, 117,* 412–414.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology*: *Human Perception and Performance, 17,* 3–27.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. In A. A. J. Marley (Ed.), *Choice, measurement, and decision: Essays in honor of R. Duncan Luce* (pp. 347–366). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Palmeri, T. J. (1997a). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception & Psychophysics, 59,* 1027–1048.

Nosofsky, R. M., & Palmeri, T. J. (1997b). 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., & 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.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

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

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 659–680.

Thomas, R. D. (1996). Separability and independence of dimensions in the same–different judgment task. *Journal of Mathematical Psychology, 40,* 318–341.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes.* New York: Cambridge University Press.

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.

Appendix A

Random-Walk Version of the Linear Decision-Boundary Model

In this model, the perceptual representation and decision boundaries are the same as in the standard version of decision-boundary theory. As is the case in the exemplar-based random-walk (exemplar-RW) model, there is a random-walk counter with initial value 0 and decision criteria set at +A and −B. Assume that stimulus i is presented. On each step of the random walk, a percept is sampled from the perceptual distribution associated with stimulus i. If the percept falls in Region A of the space, then the random walk takes a step in the direction of Boundary +A; otherwise, it takes a step in the direction of Boundary −B. The perceptual-sampling process continues until either Boundary +A or −B is reached. The probability that a stimulus gives rise to a percept that falls in Region A, p_i, is computed by using the same method as assumed in the standard version of decision-boundary theory. The prediction equations of choice probability and mean RT are then the same as already described by Nosofsky and Palmeri (1997b, Equations 14–21) for the exemplar-RW model, except with the alternative definition of p_i described above. The free parameters in this random-walk version of the linear decision-boundary model are the slope (m) and y-intercept (b) of the linear boundary, a perceptual variance parameter σ_p^2, the decision criteria +A and −B, the residual time parameter μ, and the time-scaling constant k.

Appendix B

Description of the Version of the Exemplar Model Used to Generate Figure 3

We start by reviewing the manner in which one version of the standard generalized context model was applied by Rouder and Ratcliff (2004) to their experimental paradigm. The distance between stimuli i and j is given by $d_{ij} = |i - j|$. The similarity between i and j is an exponential function of this distance, $s_{ij} = \exp(-c \cdot d_{ij})$, where c is the overall sensitivity parameter. And the probability that stimulus i is classified in Category A is given by $P(\text{Ali}) = b_A \cdot S_{iA}^{\gamma}/[b_A \cdot S_{iA}^{\gamma} + (1 - b_A) \cdot S_{iB}^{\gamma}]$, where b_A ($0 \leq b_A \leq 1$) is the response bias for Category A, γ is the response-scaling parameter, and S_{iA} denotes the summed activation of stimulus i to the Category-A exemplars. This summed activation is computed in the same manner as described in the Overview of the Formal Models section of the introduction.

The version of the exemplar model used to generate Figure 3 is the same as that just described, except that assumptions are introduced about the role of sensory and memory noise. Specifically, it is assumed that across trials, each exemplar gives rise to a distribution of sensory effects. The stimulus-i distribution is normally distributed with mean i and variance σ_s^2. Likewise, the memory representation for exemplar i is also normally distributed with mean i and variance σ_m^2. The summed activations are computed in the same manner as for the standard model, except that instead of summing the similarity of stimulus i to single-point exemplar representations, one sums the similarity of the individual stimulus-i sensory effects to each entire exemplar-based memory distribution. A predicted response probability is obtained for each sensory effect to which stimulus i gives rise. The overall predictions for stimulus i are found by integrating across the response probabilities associated with the individual sensory effects (see Nosofsky, 1997, for an application of a similar version of such a model in the domain of unidimensional absolute identification). The parameter values used to generate Figure 3 were $c = 4.459$, $b_A = .700$, $\gamma = 4.788$, $\sigma_s = 1.500$, and $\sigma_m = .413$.

(*Appendixes follow*)

Appendix C

Observed and Predicted Category-A Choice Probabilities for Each Individual Subject and Each Stimulus in Experiment 1

	Stimulus											
Subject and model	1	2	3	4	5	6	7	8	9	10	11	12
Obs (1)	.98	1.00	.92	.92	.94	.94	.12	.09	.10	.08	.01	.01
Exemplar-RW	1.00	1.00	.93	.93	.96	.94	.13	.13	.07	.11	.01	.01
Decision bound	1.00	1.00	.93	.93	.93	.93	.10	.10	.10	.09	.00	.00
Prototype-RW	1.00	1.00	.90	.98	.98	.89	.19	.05	.05	.18	.01	.01
Obs (2)	.97	.98	.74	.74	.92	.74	.13	.14	.09	.20	.02	.03
Exemplar-RW	.97	.98	.76	.83	.89	.86	.14	.14	.13	.22	.02	.02
Decision bound	1.00	1.00	.83	.86	.88	.90	.12	.14	.16	.18	.00	.00
Prototype-RW	.97	.97	.70	.88	.90	.79	.20	.09	.12	.29	.02	.02
Obs (3)	.95	.91	.93	.80	.76	.56	.24	.17	.10	.10	.03	.05
Exemplar-RW	.96	.95	.86	.83	.79	.61	.32	.20	.09	.08	.03	.02
Decision bound	1.00	.99	.90	.83	.73	.62	.28	.18	.11	.06	.00	.00
Prototype-RW	.95	.95	.82	.89	.79	.54	.40	.16	.08	.14	.03	.03
Obs (4)	1.00	1.00	.95	.87	.97	.97	.09	.12	.02	.07	.01	.00
Exemplar-RW	1.00	1.00	.95	.89	.97	.98	.04	.12	.04	.05	.01	.00
Decision bound	1.00	1.00	.97	.97	.97	.97	.05	.05	.05	.05	.00	.00
Prototype-RW	1.00	1.00	.95	1.00	1.00	.95	.12	.02	.02	.13	.00	.00
Obs (5)	.99	.99	.98	.94	.95	.91	.17	.07	.05	.07	.00	.00
Exemplar-RW	1.00	1.00	.95	.93	.96	.91	.11	.09	.03	.03	.00	.00
Decision bound	1.00	1.00	.97	.96	.94	.91	.11	.08	.06	.04	.00	.00
Prototype-RW	1.00	1.00	.93	.99	.98	.83	.20	.03	.02	.08	.00	.00
Obs (6)	1.00	1.00	.93	.77	.97	.83	.16	.20	.04	.06	.00	.00
Exemplar-RW	.99	1.00	.92	.91	.95	.92	.10	.10	.05	.06	.01	.00
Decision bound	1.00	1.00	.96	.95	.93	.92	.12	.10	.08	.06	.00	.00
Prototype-RW	1.00	1.00	.89	.97	.97	.83	.17	.03	.03	.11	.00	.00
Obs (7)	.99	.98	.86	.91	.85	.85	.19	.27	.29	.61	.05	.07
Exemplar-RW	.98	.99	.81	.88	.92	.91	.27	.26	.26	.38	.08	.08
Decision bound	1.00	1.00	.83	.87	.90	.92	.25	.30	.35	.41	.02	.02
Prototype-RW	.98	.98	.77	.91	.93	.85	.33	.20	.25	.44	.09	.09
Obs (8)	1.00	.98	.90	.77	.91	.80	.13	.06	.07	.06	.00	.00
Exemplar-RW	.99	.99	.87	.88	.92	.86	.11	.09	.05	.07	.01	.00
Decision bound	1.00	1.00	.88	.87	.86	.85	.08	.07	.07	.06	.00	.00
Prototype-RW	.99	.99	.83	.94	.93	.76	.19	.05	.04	.12	.00	.00
Obs (9)	.96	.88	.97	.88	.71	.83	.44	.16	.07	.31	.02	.05
Exemplar-RW	.99	.98	.96	.93	.79	.71	.34	.16	.15	.08	.03	.02
Decision bound	1.00	1.00	.95	.91	.84	.74	.32	.21	.12	.06	.00	.00
Prototype-RW	.99	.99	.91	.96	.91	.69	.46	.18	.10	.18	.04	.03
Obs (10)	.90	.83	.69	.79	.81	.78	.25	.29	.14	.25	.03	.05
Exemplar-RW	.95	.95	.77	.83	.80	.77	.25	.19	.23	.28	.06	.06
Decision bound	.99	.99	.78	.78	.79	.79	.24	.24	.25	.25	.01	.02
Prototype-RW	.94	.94	.71	.84	.84	.72	.32	.18	.19	.32	.07	.07
Obs (11)	.98	.97	.90	.94	.70	.94	.13	.15	.16	.07	.01	.02
Exemplar-RW	.99	.98	.92	.92	.82	.87	.13	.09	.18	.11	.02	.02
Decision bound	1.00	1.00	.92	.91	.90	.88	.16	.14	.12	.11	.00	.00
Prototype-RW	.99	.99	.85	.96	.95	.83	.20	.06	.06	.18	.01	.01
Obs (12)	.97	.97	.85	.92	.86	.93	.15	.08	.11	.31	.01	.03
Exemplar-RW	.99	.99	.86	.92	.89	.89	.12	.09	.14	.18	.01	.02
Decision bound	1.00	1.00	.87	.89	.91	.93	.10	.11	.14	.16	.00	.00
Prototype-RW	.99	.99	.78	.93	.95	.86	.19	.07	.10	.27	.02	.02
Obs (13)	1.00	1.00	.98	.98	.98	.93	.04	.04	.07	.01	.00	.00
Exemplar-RW	1.00	1.00	.99	1.00	.98	.98	.04	.02	.03	.04	.00	.00
Decision bound	1.00	1.00	.99	.99	.98	.98	.04	.04	.04	.03	.00	.00
Prototype-RW	1.00	1.00	.99	1.00	1.00	.97	.06	.00	.00	.04	.00	.00
Obs (14)	.99	.98	.95	.96	.94	.85	.20	.14	.16	.04	.02	.01
Exemplar-RW	1.00	1.00	.98	.98	.91	.85	.24	.10	.11	.07	.01	.01
Decision bound	1.00	1.00	.97	.95	.92	.88	.22	.15	.10	.06	.00	.00
Prototype-RW	1.00	1.00	.94	.98	.96	.76	.32	.07	.04	.10	.01	.01
Obs (15)	1.00	.98	.92	.95	.92	.85	.19	.18	.15	.11	.03	.01
Exemplar-RW	.99	.99	.93	.95	.89	.81	.22	.11	.11	.11	.01	.01
Decision bound	1.00	1.00	.96	.93	.90	.85	.22	.16	.11	.07	.00	.00
Prototype-RW	.99	.99	.88	.96	.93	.75	.29	.09	.06	.15	.01	.01
Obs (16)	.99	.94	.94	.88	.88	.80	.11	.24	.19	.17	.00	.04
Exemplar-RW	.99	.98	.88	.91	.86	.86	.13	.10	.16	.15	.02	.02
Decision bound	1.00	1.00	.88	.88	.88	.88	.18	.18	.17	.17	.00	.00
Prototype-RW	.99	.99	.82	.93	.92	.77	.24	.08	.07	.19	.01	.01

Note. Subjects 1–8 participated in Condition 4/8; Subjects 9–16 participated in Condition 5/9. Obs (n) = observed data for Subject n, Exemplar-RW = exemplar-based random-walk model, Decision bound = linear decision-boundary model, Prototype-RW = prototype-based random-walk model.

Appendix D

Observed and Predicted Mean Response Times (in Milliseconds) for Each Individual Subject and Each Stimulus in Experiment 1

	Stimulus											
Subject and model	1	2	3	4	5	6	7	8	9	10	11	12
Obs (1)	336	336	362	369	364	364	380	384	380	376	351	352
Exemplar-RW	339	337	372	372	361	368	383	383	372	379	348	346
Decision bound	340	340	365	365	366	366	382	381	380	379	340	340
Prototype-RW	342	342	375	355	356	376	387	370	369	386	355	355
Obs (2)	425	433	474	480	449	549	475	463	481	545	423	431
Exemplar-RW	430	425	496	484	468	476	477	475	473	492	425	426
Decision bound	428	428	492	476	464	455	461	472	486	505	428	428
Prototype-RW	429	428	508	475	466	495	493	463	472	506	426	426
Obs (3)	576	618	611	625	667	685	645	664	651	662	560	586
Exemplar-RW	596	602	637	645	653	674	668	650	620	615	586	575
Decision bound	582	592	634	647	661	676	662	648	635	623	583	574
Prototype-RW	596	598	647	627	653	678	676	642	614	636	585	583
Obs (4)	558	549	586	656	592	557	647	637	608	620	548	549
Exemplar-RW	550	541	609	643	586	585	607	648	600	612	559	555
Decision bound	551	551	597	598	598	598	623	622	622	621	551	551
Prototype-RW	558	558	609	577	577	607	629	595	596	630	570	570
Obs (5)	609	616	661	727	693	701	744	734	675	703	577	572
Exemplar-RW	594	594	691	714	686	735	744	732	667	671	596	584
Decision bound	589	591	673	687	702	720	733	714	697	682	592	590
Prototype-RW	601	602	703	646	664	752	761	674	656	713	609	608
Obs (6)	586	594	817	817	627	934	854	784	656	780	578	568
Exemplar-RW	578	571	749	764	706	751	772	774	700	727	581	568
Decision bound	581	581	680	694	709	727	764	742	723	706	583	582
Prototype-RW	578	579	800	678	696	844	844	696	678	800	579	579
Obs (7)	588	601	604	642	661	682	661	672	650	723	596	627
Exemplar-RW	597	594	650	639	626	631	661	660	660	669	627	628
Decision bound	597	594	647	642	637	632	657	663	669	676	611	615
Prototype-RW	600	599	654	630	623	642	664	651	657	669	629	629
Obs (8)	528	528	582	630	581	649	625	581	591	609	506	515
Exemplar-RW	529	527	617	612	596	622	611	598	573	590	514	510
Decision bound	521	523	604	606	609	611	595	593	591	589	515	513
Prototype-RW	530	530	629	580	590	647	634	573	564	613	517	516
Obs (9)	406	443	402	443	485	460	482	473	464	491	425	422
Exemplar-RW	408	420	433	443	475	486	489	467	465	448	428	422
Decision bound	418	420	442	452	464	479	488	470	457	446	422	419
Prototype-RW	414	415	449	433	449	479	486	467	452	467	433	432
Obs (10)	525	540	651	565	577	590	623	615	594	640	555	565
Exemplar-RW	537	539	605	591	598	605	608	596	605	612	545	547
Decision bound	543	543	597	597	597	597	601	602	602	602	546	546
Prototype-RW	538	538	618	587	586	617	622	595	595	623	547	548
Obs (11)	453	467	518	514	494	514	528	531	527	492	458	449
Exemplar-RW	456	463	495	494	521	511	510	501	522	506	464	467
Decision bound	457	457	497	502	507	513	532	524	517	510	457	457
Prototype-RW	463	463	521	491	493	525	531	500	498	527	469	469
Obs (12)	568	621	678	638	723	719	742	654	695	873	579	616
Exemplar-RW	581	583	705	670	688	693	697	677	709	725	590	597
Decision bound	594	591	688	682	677	671	679	685	691	697	595	597
Prototype-RW	584	584	738	663	649	711	730	671	685	754	603	604
Obs (13)	459	473	514	515	510	547	557	556	571	570	463	470
Exemplar-RW	455	458	521	504	532	544	567	536	559	563	474	476
Decision bound	465	465	517	519	521	523	565	561	558	555	465	465
Prototype-RW	469	469	540	494	498	558	589	519	514	569	481	481
Obs (14)	528	546	539	558	579	641	620	644	619	548	532	525
Exemplar-RW	511	520	545	551	594	612	633	598	601	583	544	540
Decision bound	527	527	553	564	580	604	659	619	591	571	528	528
Prototype-RW	532	533	576	553	569	614	622	582	566	590	543	542
Obs (15)	654	819	891	847	896	1,075	1,170	933	842	788	682	642
Exemplar-RW	643	662	821	800	885	953	974	880	880	879	693	687
Decision bound	666	669	785	825	877	946	1,054	959	888	833	673	669
Prototype-RW	668	670	889	781	825	975	994	855	811	916	696	693
Obs (16)	584	590	629	722	639	729	651	749	697	685	548	615
Exemplar-RW	568	574	661	644	674	671	669	651	681	678	575	581
Decision bound	571	571	654	655	655	656	685	684	684	683	574	574
Prototype-RW	580	581	673	629	637	686	688	640	633	676	584	583

Note. Subjects 1–8 participated in Condition 4/8; Subjects 9–16 participated in Condition 5/9. Obs (*n*) = observed data for Subject *n;* Exemplar-RW = exemplar-based random-walk model; Decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model.

Appendix E

Observed and Predicted Category-A Choice Probabilities for Each Individual Subject and Each Stimulus in Experiment 2

	Stimulus											
Subject and model	1	2	3	4	5	6	7	8	9	10	11	12
Obs (1)	.99	1.00	.94	.82	.90	.94	.09	.27	.06	.17	.03	.00
Exemplar-RW	.99	.99	.95	.86	.95	.93	.10	.18	.05	.05	.01	.01
Decision bound	1.00	1.00	.93	.92	.91	.90	.14	.12	.11	.10	.00	.00
Prototype-RW	1.00	1.00	.87	.97	.97	.87	.20	.06	.06	.19	.01	.01
Obs (2)	.99	1.00	.99	.95	1.00	.99	.03	.12	.01	.00	.00	.00
Exemplar-RW	1.00	1.00	.98	.96	.99	.99	.01	.02	.00	.00	.00	.00
Decision bound	1.00	1.00	.98	.99	1.00	1.00	.00	.00	.00	.00	.00	.00
Prototype-RW	1.00	1.00	.98	1.00	1.00	.98	.00	.00	.00	.01	.00	.00
Obs (3)	1.00	1.00	.83	.84	.94	.98	.06	.10	.03	.01	.01	.00
Exemplar-RW	.99	1.00	.93	.88	.96	.94	.04	.08	.02	.03	.00	.00
Decision bound	1.00	1.00	.84	.89	.93	.95	.00	.01	.01	.02	.00	.00
Prototype-RW	1.00	1.00	.93	.99	.99	.94	.02	.00	.00	.03	.00	.00
Obs (4)	1.00	1.00	.90	.95	.98	.98	.01	.09	.02	.03	.00	.00
Exemplar-RW	1.00	1.00	.97	.93	.98	.97	.03	.05	.01	.01	.00	.00
Decision bound	1.00	1.00	.92	.95	.97	.99	.00	.00	.00	.00	.00	.00
Prototype-RW	1.00	1.00	.90	.99	.99	.95	.00	.00	.00	.01	.00	.00
Obs (5)	.99	.98	.90	.77	.96	.83	.12	.23	.05	.06	.00	.01
Exemplar-RW	.99	.99	.93	.86	.95	.93	.09	.15	.05	.06	.01	.01
Decision bound	1.00	1.00	.94	.93	.92	.90	.13	.11	.09	.07	.00	.00
Prototype-RW	.99	.99	.86	.96	.95	.81	.20	.05	.04	.14	.01	.00
Obs (6)	.99	1.00	.93	.94	.95	.92	.08	.03	.06	.03	.00	.01
Exemplar-RW	1.00	1.00	.96	.98	.95	.92	.08	.04	.05	.06	.00	.00
Decision bound	1.00	1.00	.96	.95	.95	.94	.06	.05	.05	.04	.00	.00
Prototype-RW	1.00	1.00	.93	.99	.98	.88	.12	.02	.01	.07	.00	.00
Obs (7)	1.00	1.00	.99	.92	.92	.94	.03	.05	.03	.17	.00	.00
Exemplar-RW	1.00	1.00	.97	.98	.95	.96	.03	.02	.06	.05	.00	.00
Decision bound	1.00	1.00	.97	.97	.97	.97	.04	.04	.05	.05	.00	.00
Prototype-RW	1.00	1.00	.96	1.00	1.00	.97	.07	.01	.01	.09	.00	.00
Obs (8)	.99	.98	.98	.95	.90	.99	.05	.10	.02	.06	.01	.00
Exemplar-RW	1.00	1.00	.98	.99	.95	.96	.01	.00	.01	.01	.00	.00
Decision bound	1.00	1.00	1.00	1.00	1.00	.98	.01	.00	.00	.00	.00	.00
Prototype-RW	1.00	1.00	.98	1.00	1.00	.97	.01	.00	.00	.00	.00	.00
Obs (9)	1.00	1.00	.96	.93	.88	.98	.01	.02	.05	.01	.00	.01
Exemplar-RW	1.00	1.00	1.00	1.00	1.00	1.00	.00	.00	.00	.00	.00	.00
Decision bound	1.00	1.00	.98	.97	.97	.97	.03	.03	.02	.02	.00	.00
Prototype-RW	1.00	1.00	1.00	1.00	1.00	.99	.02	.00	.00	.01	.00	.00
Obs (10)	1.00	.96	.93	.94	.90	.95	.11	.13	.10	.05	.00	.00
Exemplar-RW	1.00	.99	.96	.97	.89	.92	.09	.06	.12	.07	.01	.01
Decision bound	1.00	1.00	.95	.94	.93	.92	.12	.11	.10	.08	.00	.00
Prototype-RW	1.00	1.00	.93	.99	.99	.90	.18	.04	.03	.13	.00	.00

Note. Subjects 1–5 participated in Condition 4/8; Subjects 6–10 participated in Condition 5/9. Obs (*n*) = observed data for Subject *n;* Exemplar-RW = exemplar-based random-walk model; Decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model.

Appendix F

Observed and Predicted Mean Response Times (in Milliseconds) for Each Individual Subject and Each Stimulus in Experiment 2

	Stimulus											
Subject and model	1	2	3	4	5	6	7	8	9	10	11	12
Obs (1)	599	634	693	933	735	726	856	954	713	794	631	614
Exemplar-RW	611	601	733	848	726	762	814	892	749	739	652	621
Decision bound	614	615	731	743	755	769	813	796	780	765	616	615
Prototype-RW	622	622	778	694	696	783	813	731	729	809	649	649
Obs (2)	544	522	630	656	598	596	612	686	549	611	477	482
Exemplar-RW	518	509	636	694	593	607	590	647	559	577	492	485
Decision bound	522	518	628	624	619	614	566	570	574	578	483	486
Prototype-RW	522	522	663	567	565	648	601	536	538	614	504	504
Obs (3)	607	597	668	711	668	685	660	670	657	628	578	585
Exemplar-RW	610	603	680	709	661	672	656	685	634	642	587	580
Decision bound	602	598	690	684	678	672	647	652	657	663	582	586
Prototype-RW	615	615	695	645	644	692	666	625	626	668	603	603
Obs (4)	596	603	720	673	664	679	636	748	611	613	539	573
Exemplar-RW	591	586	666	711	652	674	656	689	618	621	570	561
Decision bound	596	593	685	681	678	674	629	632	635	638	560	562
Prototype-RW	609	609	719	648	641	691	634	602	606	656	585	585
Obs (5)	600	613	916	1019	676	880	905	970	724	819	604	674
Exemplar-RW	623	604	803	897	766	797	831	910	766	777	631	604
Decision bound	614	615	753	771	792	816	857	829	804	782	616	615
Prototype-RW	628	629	877	749	770	921	923	772	752	880	631	630
Obs (6)	868	904	1,042	1,070	1,035	1,250	1,133	1,109	1,125	1,122	878	959
Exemplar-RW	892	903	1,073	1,033	1,104	1,150	1,152	1,070	1,105	1,116	912	914
Decision bound	900	902	1,085	1,090	1,095	1,101	1,101	1,095	1,090	1,085	902	900
Prototype-RW	936	937	1,131	1,018	1,035	1,184	1,185	1,036	1,019	1,133	938	937
Obs (7)	567	589	625	692	667	681	623	704	684	729	558	586
Exemplar-RW	573	578	659	640	683	671	661	643	688	677	575	581
Decision bound	572	572	666	665	665	665	671	671	672	672	574	574
Prototype-RW	589	589	668	616	614	659	681	631	633	690	599	599
Obs (8)	585	645	617	675	757	687	651	719	612	678	559	540
Exemplar-RW	594	602	670	662	727	715	641	616	654	630	562	564
Decision bound	565	588	622	649	678	709	709	678	649	621	587	564
Prototype-RW	599	600	701	635	638	721	673	610	608	657	583	583
Obs (9)	648	695	728	804	905	807	821	862	810	848	658	723
Exemplar-RW	668	677	772	756	844	852	856	791	850	821	686	688
Decision bound	676	677	819	821	823	825	823	821	819	817	676	675
Prototype-RW	691	691	799	727	731	822	887	768	763	858	713	712
Obs (10)	553	620	604	680	697	630	676	715	684	659	569	611
Exemplar-RW	567	578	633	630	687	669	679	657	697	667	590	594
Decision bound	577	577	643	650	658	666	700	688	678	669	578	578
Prototype-RW	594	594	659	621	625	671	689	644	639	678	607	607

Note. Subjects 1–5 participated in Condition 4/8; Subjects 6–10 participated in Condition 5/9. Obs (*n*) = observed data for Subject *n;* Exemplar-RW = exemplar-based random-walk model; Decision bound = linear decision-boundary model; Prototype-RW = prototype-based random-walk model.

Received May 13, 2003
Revision received November 23, 2004
Accepted November 29, 2004 ■