Journal of Experimental Psychology: General
2004, Vol. 133, No. 1, 63–82

0096-3445/04/$12.00 DOI: 10.1037/0096-3445.133.1.63

Comparing Categorization Models

Jeffrey N. Rouder
University of Missouri—Columbia

Roger Ratcliff
The Ohio State University

Four experiments are presented that competitively test rule- and exemplar-based models of human categorization behavior. Participants classified stimuli that varied on a unidimensional axis into 2 categories. The stimuli did not consistently belong to a category; instead, they were probabilistically assigned. By manipulating these assignment probabilities, it was possible to produce stimuli for which exemplar- and rule-based explanations made qualitatively different predictions. F. G. Ashby and J. T. Townsend's (1986) rule-based general recognition theory provided a better account of the data than R. M. Nosofsky's (1986) exemplar-based generalized context model in conditions in which the to-be-classified stimuli were relatively confusable. However, generalized context model provided a better account when the stimuli were relatively few and distinct. These findings are consistent with multiple process accounts of categorization and demonstrate that stimulus confusion is a determining factor as to which process mediates categorization.

In this article we present an empirical paradigm to test different theories of categorization behavior. One theory we test is the exemplar-based theory in which categories are represented by sets of stored exemplars. Category membership of a stimulus is determined by similarity of the stimulus to these exemplars (e.g., Medin & Schaffer, 1978; Nosofsky, 1986, 1987, 1991). An exemplar-based process relies on retrieval of specific trace-based information without further abstraction; for example, a person is judged as "tall" if he or she is similar in height to others who are considered "tall." The other theory we test is rule-based or decision-bound theory. Decisions are based on an abstracted rule. The relevant space is segmented into regions by bounds, and each region is assigned to a specific category (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1992, 1993; Ashby & Perrin, 1988; Ashby & Townsend, 1986; Trabasso & Bower, 1968). For example, a person might be considered tall if he or she is perceived as being over 6 ft (1.83 m). The essence of a rule-based process is that processing is based on greatly simplified abstractions or rules but not on the specific trace-based information itself.

Both rule- and exemplar-based theories have gained a large degree of support in the experimental literature. There are both exemplar-based models (e.g., generalized context model; Nosofsky, 1986) and rule-based models (e.g., general recognition theory; Ashby & Gott, 1988; Ashby & Townsend, 1986) that can explain a wide array of behavioral data across several domains. Despite the many attempts to discriminate between these two explanations, there have been few decisive tests. Across several paradigms and domains, rule- and exemplar-based predictions often mimic each other.

Roughly speaking, there are two main experimental paradigms for testing the above theories of categorization. One is a probabilistic assignment paradigm: Stimuli do not always belong to the same category. Instead, a stimulus is probabilistically assigned to categories by the experimenter. Probabilistic assignment is meant to capture the uncertainty in many real-life decision-making contexts. For example, consider the problem of interpreting imperfect medical tests. The test result may indicate the presence of a disease, but there is some chance of a false positive. Examples of the use of probabilistic assignment in categorization tasks include the studies of Ashby and Gott (1988), Espinoza-Varas and Watson (1994), Kalish and Kruschke (1997), McKinley and Nosofsky (1995), Ratcliff, Van Zandt, and McKoon (1999), and Thomas (1998). The other main paradigms for testing categorization models are transfer paradigms. In these paradigms, participants first learn a set of stimulus–category assignments. Afterward, they are given novel stimuli to categorize, and the pattern of responses to these novel stimuli serves as the main dependent measure. Examples of the use of transfer include the studies of Erickson and Kruschke (1998), Medin and Schaffer (1978), and Nosofsky, Palmeri, and McKinley (1994).

Both transfer and probabilistic assignment paradigms offer unique advantages and disadvantages. In this article, we are concerned with probabilistic assignment. We first show that the main drawback with probabilistic assignment is that in many paradigms, rule- and exemplar-based theories yield similar predictions—that is, they mimic each other. We then present a novel method of assigning categories to stimuli that mitigates this problem and provides for a differential test of the theories.

Jeffrey N. Rouder, Department of Psychological Sciences, University of Missouri—Columbia; Roger Ratcliff, Psychology Department, The Ohio State University.

This research was supported by National Science Foundation Grant BCS-9817561 and University of Missouri Research Board Grant 00-77 to Jeffrey N. Rouder, National Institute of Mental Health Grants R37-MH44640 and K05-MH01891 to Roger Ratcliff, and National Institute on Deafness and Other Communication Disorders Grant R01-DC01240 to Gail M. McKoon. We are grateful to Tonya Cottrell, Leslie Henry, and Laura White for assistance in running the reported experiments.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211. E-mail: jeff@banta.psyc.missouri.edu

We report the results of four categorization experiments with simple, unidimensional stimuli. The results point to *stimulus confusion* as a salient variable. Stimulus confusion is the inability to identify stimuli in an absolute identification task. When stimuli are confusable, the results are consistent with rule-based models and inconsistent with exemplar models. When the stimuli are clear and distinct, however, the pattern favors an exemplar-based interpretation. The finding that both exemplar- and rule-based processes may be used in the same task is compatible with recent work in multiple-process and multiple-system frameworks (e.g., Ashby & Ell, 2002; Erickson & Kruschke, 1998; Nosofsky et al., 1994). The results provide guidance as to the type of information that determines which system or process will mediate categorization.

Perhaps the best way to illustrate exemplar- and rule-based theories and their similar predictions is with a simple experiment. Ratcliff and Rouder (1998) asked participants to classify squares of varying luminance into one of two categories (denoted Category A and Category B). For each trial, luminance of the stimulus was chosen from one of the two categories. The stimuli from a given category were selected from a normal distribution on a luminance scale. The two categories had the same standard deviation but different means (see Figure 1A). Feedback was given after each trial to tell the participants whether the decision had correctly indicated the category from which the stimulus had been drawn. Because the categories overlapped substantially, participants could not be highly accurate. For example, a stimulus of moderate luminance might have come from Category A on one trial and from Category B on another. But overall, dark stimuli tended to come from Category A and light stimuli from Category B.

According to rule-based explanations, participants partition the luminance dimension into regions and associate these regions with certain categories. The dotted line in Figure 1A denotes a possible criterion placement that partitions the luminance domain into two regions. If the participant perceives the stimulus as having luminance greater than the criterion, the participant produces the Category B response; otherwise the participant produces the Category A response. A participant may not perfectly perceive the luminance of a stimulus, and perceptual noise can explain gradations in the participant's empirical response proportions. Alternatively, there may be trial-by-trial variability in the placement of the criteria. This variability will also lead to graded response proportions (Wickelgren, 1968). Ashby and Maddox (1992; see Maddox & Ashby, 1993) showed that in the presence of perceptual noise, rule-based explanations have the property that response probability is dependent on the psychological distance between the decision criterion and the stimulus. Overall, the rule-based explanation predicts that as luminance increases, the probability of a Category A response decreases. This prediction is not predicated on the exact placement of the criterion; it holds for all criteria that partition the luminance into two regions, as in Figure 1A.

According to exemplar-based explanations, participants make Category A responses when they perceive the stimulus as similar to Category A exemplars and dissimilar from Category B exemplars. Figure 1B shows the relative proportion of Category A exemplars as a function of luminance for Ratcliff and Rouder's (1998) design. There are relatively more Category A exemplars for dark stimuli and relatively more Category B exemplars for light stimuli. As noted by Ashby and Alfonso-Reese (1995), exemplar-based categorization response probability predictions tend to follow these relative proportions. For Ratcliff and Rouder's design, exemplar-based explanations make the same qualitative predictions as the rule-based explanations: The probability of Category A response decreases with increasing luminance.

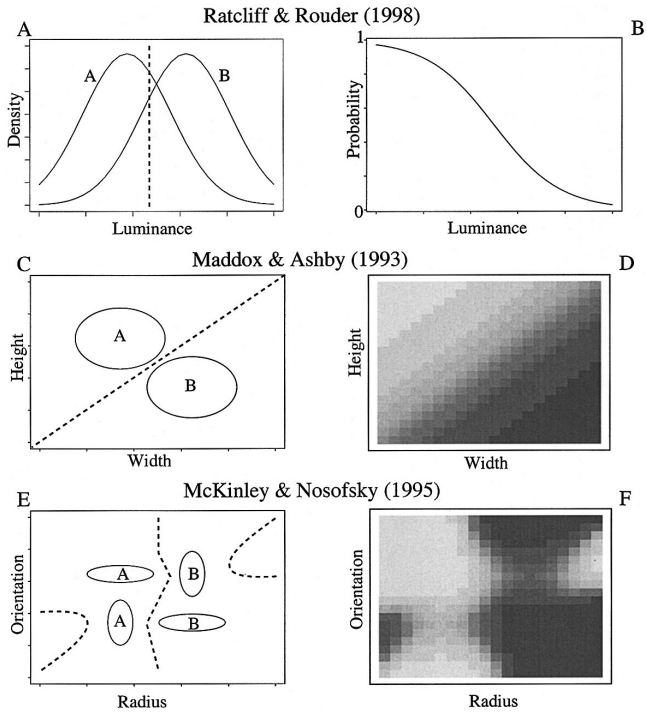

Figure 1. A: Category structure and decision bound for Ratcliff and Rouder's (1998) univariate design. B: Relative proportion of Category A exemplars for Ratcliff and Rouder's design. C: Category structure and decision bound for Maddox and Ashby's (1993) design. Categories were composed of bivariate normal distributions, which are represented as ellipses. D: Relative proportion of Category A exemplars for Maddox and Ashby's design. Lighter areas correspond to higher proportions. E and F: Category structure, decision bounds, and relative proportion of Category A exemplars for McKinley and Nosofsky's (1995) design. Categories were composed of mixtures of bivariate normal distributions; each component is represented with an ellipse. Panel A is from "Modeling Response Times for Decisions Between Two Choices," by R. Ratcliff and J. N. Rouder, 1998, *Psychological Science, 9,* p. 350. Copyright 1998 by Blackwell Publishing. Adapted with permission. Panel C is from "Comparing Decision Bound and Exemplar Models of Categorization," by W. T. Maddox and F. G. Ashby, 1993, *Perception & Psychophysics, 53,* p. 55. Copyright 1993 by the Psychonomic Society. Adapted with permission. Panel E is from "Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures," by S. C. McKinley and R. M. Nosofsky, 1995, *Journal of Experimental Psychology: Human Perception and Performance, 21,* p. 135. Copyright 1995 by the American Psychological Association. Adapted with permission of the author.

This covariation of decision regions and exemplars is prevalent in designs in complex, multidimensional stimuli as well. We review two different multivariate designs (shown in Figures 1C and 1E). Maddox and Ashby (1993) had participants classify rectangles that varied in two dimensions: height and width. Category membership was determined by assuming that each category was distributed over the height and width of the rectangles as an uncorrelated bivariate normal (Figure 1C). McKinley and Nosof-

sky (1995) had participants classify circles with an embedded diameter; both the radius of the circle and the orientation of the embedded diameter varied. The categories were distributed as a mixture of normal distributions. Optimal decision bounds for both designs are indicated with dotted lines. The relative proportion of Category A exemplars is shown in Figures 1D and 1F, respectively. For both designs, the regions with a high proportion of Category A exemplars tend to be those that correspond to Category A according to the rule-based partitions. Given that exemplar- and rule-based explanations make the same qualitative predictions in the above designs, it is not surprising that the data yield roughly the same level of support for both types of explanations (see Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). Thus, in the three designs of Figure 1, the patterns of results predicted from rule- and exemplar-based explanations generally mimic each other.

We propose a novel design for probabilistically assigning stimuli to categories. The main goal is to provide a differential test of rule- and exemplar-based explanations, that is, to avoid the mimicking problems in the above designs. In our experiments, categories were distributed in a complex fashion. Figure 2A shows the assignment for Experiment 1. The stimuli were unidimensional—they were squares that varied in luminance. Stimuli that were extreme, either extremely dark or extremely light, were assigned by the experimenter slightly more often to Category A than to Category B. Moderately dark stimuli were always assigned to Category A, and moderately light stimuli were always assigned to Category B. The most reasonable rule-based approach was to partition the space into three regions with two bounds. These bounds, along with response probability predictions for this arrangement, are shown in Figure 2B. Exemplar models are both complex and flexible, but for this design, response probabilities tend to follow category assignment probabilities (see Ashby & Alphonso-Reese, 1995). The predictions shown in Figure 2C are typical for reasonable parameters in the generalized context model. The critical stimuli in the design are the dark ones. According to rule-based models, the Category A response probability should increase monotonically with decreasing luminance. According to the exemplar-based models, the Category A response probability should decrease with decreasing luminance.

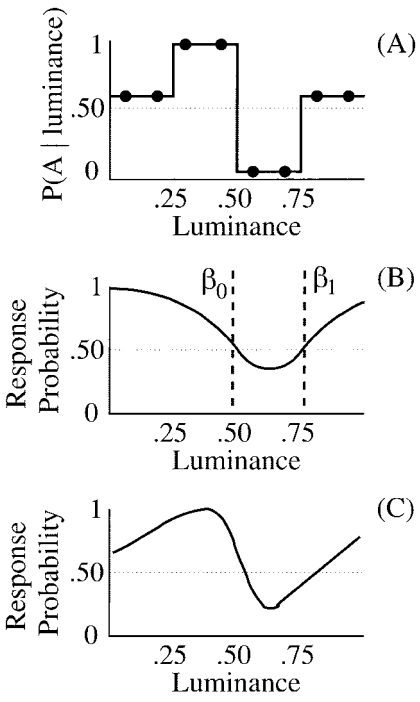

Figure 2. A: Category assignment probabilities as a function of luminance. B: Typical rule-based predictions. C: Typical exemplar-based predictions.

Unlike previous designs, our relatively simple design, shown in Figure 2, has the ability to distinguish between exemplar- and rule-based explanations. The use of unidimensional stimuli restricts the complexity of exemplar- and rule-based models. Exemplar-based models of multidimensional stimuli often have adjustable attentional weights on the dimensions. By selectively tuning these weights, the model becomes very flexible (see, e.g., Kruschke, 1992; Nosofsky & Johansen, 2000). But with unidimensional stimuli, attention is set to the single dimension rather than tuned. Likewise, rule-based models are also constrained in one dimension; the set of plausible rules is greatly simplified.

Generalized Context Model (GCM)

GCM is a well-known exemplar model that has had much success in accounting for data from categorization experiments (Nosofsky, 1986; Nosofsky & Johansen, 2000). According to GCM, participants make category decisions by comparing the similarity of the presented stimulus to each of several stored category exemplars. In general, both the stimuli and exemplars are represented in a multidimensional space. But in Experiment 1, the stimuli varied on a single dimension: luminance. Therefore, we implement a version of the GCM in which stimuli are represented on a unidimensional line termed *perceived luminance.*

In GCM performance is governed by the similarity of the perceived stimulus to stored exemplars. Each exemplar is represented as a point on the perceived-luminance line. Let x_j denote the perceived luminance of the jth exemplar, and let y denote the perceived luminance of the presented stimulus. The distance between exemplar j and the stimulus on the perceived-luminance line is simply $d_j = |x_j - y|$. In GCM, the similarity of a stimulus to an exemplar j is related to this distance (see also Shepard, 1957, 1987) as

$$s_j = \exp(-\lambda d_j^k), \quad \lambda > 0. \tag{1}$$

The exponent k describes the shape of the similarity gradient. Smaller values of k correspond to a sharper decrease in similarity with distance, whereas higher values correspond to a shallowed decrease. Two values of k are traditionally chosen: $k = 1$ corresponds to an exponential similarity gradient, and $k = 2$ corresponds to a Gaussian similarity gradient. In the analyses presented here, we separately fit GCM with both the exponential ($k = 1$) and Gaussian ($k = 2$) gradients. Although the value of k is critically important in determining the categorization of novel stimuli far in distance from previous exemplars (Nosofsky & Johansen, 2000), it is less critical in the probabilistic assignment paradigms in which each stimulus is nearby to several exemplars. In our analyses, varying k only marginally affected model fits. The parameter λ is referred to as the *sensitivity,* and it describes how similarity lin-

early scales with distance. If λ is large, then similarity decreases rapidly with increasing distance. However, if λ is small, similarity decreases slowly with increasing distance.

The pivotal quantity in determining the response probabilities is the similarity of the stimulus to all of the exemplars. Each exemplar is associated with one or the other category. Let $\mathcal{A}$ be all of the exemplars associated with Category A, and let $\mathcal{B}$ be all of the exemplars associated with Category B. In GCM, the activation of a category is the sum of similarities of the stimulus to exemplars associated with that category. The activation of Category A to the stimulus is denoted by a and given by

$$a = \sum_{j \in \mathcal{A}} s_j. \tag{2}$$

Likewise, the activation of Category B is denoted by b and given by

$$b = \sum_{j \in \mathcal{B}} s_j. \tag{3}$$

Response probabilities are based on the category activations. Let $P(A)$ denote the probability that the participant places the stimulus into Category A:

$$P(A) = \frac{a^\gamma}{a^\gamma + b^\gamma + \phi}, \quad \gamma > 0. \tag{4}$$

Parameter ϕ is a response bias. If ϕ is positive, then there is a response bias toward Category A. If ϕ is negative, then there is a response bias toward Category B. The parameter γ is the consistency of response. If γ is large, responses tend to be consistent; stimuli are classified into Category A if $a > b$ and into Category B if $b > a$. As γ decreases, responses are less consistent; participants sometimes choose Category B even when $a > b$. Parameter γ was not part of GCM in its original formulation (e.g., Nosofsky, 1986) but was added by Ashby and Maddox (1993) for generality.

In GCM, each trial generates a new exemplar (e.g., Nosofsky, 1986). The feedback on the trial dictates the category membership of the new exemplar, and the new exemplar's position in psychological space is that of the stimulus. In our experiments, we used eight different luminance levels, and hence all of the exemplars fall on one of eight points on the perceived luminance dimension. We denote the category assignment probability, the probability that the experimenter assigns a stimulus with luminance i into Category A as π_i. The proportion of exemplars for Category A at a perceived luminance depends on the corresponding category assignment probability. Although this proportion may vary, as the number of trials increases this proportion converges to the category assignment probability. The category activations can be expressed in terms of the category assignment probabilities (π). For luminance level i, the number of Category A exemplars is approximately $N\pi_i$, and each of these exemplars has a similarity to the current stimulus given by s_i. Hence, in the asymptotic limit of many trials,

$$a = \sum_i N\pi_i s_i \tag{5a}$$

and

$$b = \sum_i N(1 - \pi_i)s_i. \tag{5b}$$

The model specified by Equations 1, 4, 5a, and 5b was fit in the following experiments. The three free parameters were λ, γ, and ϕ.

In subsequent exemplar models (e.g., Kruschke, 1992), exemplars could be forgotten. Although the equations were derived for the case where exemplars are perfectly retained in memory, they still apply for this version as well, providing that forgetting occurs at an equal rate for all exemplars and exemplars are presented randomly.

General Recognition Theory (GRT)

GRT is based on the assumption that people use decision rules to solve categorization problems. In GRT, perception is assumed to be inherently variable. For example, the perceived luminance of a stimulus is assumed to vary from trial to trial, even for the same stimulus. In general, the perceptual effects of stimuli in GRT can be represented in a multidimensional psychological space. For the squares of varying luminance in Experiment 1, the appropriate psychological space is the perceived luminance line. Participants use decision criteria to partition the perceived luminance line into line segments. Each segment is associated with a category, and the categorization response on a given trial depends on which segment contains the stimulus's perceived luminance.

We implemented GRT as a three-parameter model. In GRT, the perceived luminance line is partitioned into decision regions. For the current category structure (Figure 2A), it is reasonable to assume that participants set two boundaries. If the perceived luminance is between the boundaries, a Category B response is produced; otherwise a Category A response is produced. The positions of the boundaries are free parameters denoted as β_0 and β_1, where $\beta_0 < \beta_1$ (see Figure 2B). The perceived luminance of a stimulus with luminance i is assumed to be distributed as a normal with mean μ_i and variance σ_i^2. Let f denote the probability density of this normal. Then, the probability of Category A and Category B responses—$P(A)$ and $P(B)$, respectively—to a stimulus of luminance i is the integral of the density in the appropriate regions, for example,

$$P(A) = \int_{-\infty}^{\beta_0} f(x, \mu_i, \sigma_i^2)dx + \int_{\beta_1}^{\infty} f(x, \mu_i, \sigma_i^2)dx \tag{6a}$$

and

$$P(B) = \int_{\beta_0}^{\beta_1} f(x, \mu_i, \sigma_i^2)dx. \tag{6b}$$

Experiment 1

Our goal is to competitively test the specified GCM and GRT models with the design depicted in Figure 2. In Experiment 1, participants classified squares that varied in luminance. The squares were assigned to categories as depicted in Figure 2A. The question, then, is whether the pattern of responses is better fit by GRT (Figure 2B) or GCM (Figure 2C).

Method

Participants

Seven Northwestern University students participated in the experiment. They were compensated \$6 per session.

Stimuli

The stimuli were squares consisting of gray, white, and black pixels. These squares were 64 pixels in height and width and contained 4,096 pixels. Of the 4,096 pixels, 2,896 were gray. The remaining 1,200 pixels were either black or white. The numbers of black and white pixels were manipulated, but the total number of black and white pixels was always 1,200. The arrangement of black, white, and gray pixels within the square was random.

Apparatus

The experiment was conducted on PCs with 14-in. (35.56-cm) monitors programmed to a 640 × 480 square-pixel mode.

Design

The main independent variable manipulated was the proportion of the 1,200 nongray pixels that were white. The proportion was manipulated through eight levels: from .0625 (75 white pixels and 1,125 black pixels) to .9375 (1,125 white pixels and 75 black pixels) in increments of .125. Stimuli with proportions less than .5 appear dark, and those with proportions greater than .5 appear light. As shorthand, we can refer to the proportion as the *luminance* (*l*) of a stimulus.

Procedure

Each trial began with the presentation of a large gray background of 320 × 200 square pixels. Then, 500 ms later, the square of black, gray, and white pixels was placed in the center of the gray background. The square remained present until 400 ms after the participant had responded. Participants pressed the *z* key to indicate that the square belonged in Category A and the / key to indicate that the square belonged in Category B. Category assignment probability is depicted in Figure 2A. Very black and very white stimuli (e.g., stimuli with extreme proportions) were assigned to Category A with a probability of $\pi = .6$. Moderately black stimuli were always assigned to Category A ($\pi = 1.0$), and moderately white stimuli were always assigned to Category B ($\pi = 0$). If the participant's response did not match the category assignment, an error message was presented after the participant's response, and it remained on the monitor for 400 ms. A block consisted of 96 such trials. After each block, the participant was given a short break. There were 10 blocks in a session, and each session took about 35 min to complete. Participants were tested until the response probabilities were fairly stable across three consecutive sessions; this took between four and six sessions. Stability was assessed by visual inspection of the day-to-day plots or response proportions.

Instructions

Participants were instructed that category assignment was a function of the number of black and white pixels but not the arrangement of black, white, and gray pixels. They were told to learn which stimulus belonged to which category by paying careful attention to the feedback. They were also informed that the feedback was probabilistic in nature, and that they could not avoid receiving error messages on some trials. They were instructed to try to keep errors to a minimum. To explain probabilistic feedback, we provided participants with the following example from reading medical x-rays: A doctor sees a suspicious spot on an x-ray and orders a biopsy. The biopsy reveals a malignancy. In this case the doctor has correctly classified the x-ray as "problematic." The same doctor sees another x-ray from a different patient with the same type of suspicious spot. But the biopsy on this second patient reveals no malignancy; the spot is due to naturally occurring variation in tissue density. Here the doctor has classified the same stimulus (the suspicious spot) as "problematic" but is in error as there is no malignancy. Because life is variable, a response in a situation at one time is correct but the same response at another time is wrong. Participants were told that there were "tendencies"; certain luminances were more likely to belong to one category than the other. Participants were also told that the order of presentation was random and that only the luminance on the current trial determined the probability that the stimulus belongs to one or the other category. If a participant's data for the first session showed a monotonic pattern in which dark stimuli were placed into Category A and light stimuli into Category B, the participant was instructed that there was a better solution and was asked to find it.

Results

One of the 7 participants withdrew after three sessions. The data are analyzed for each of the remaining 6 participants individually and are from the last three sessions. Response probability, the probability that the participant classified a stimulus into Category A, is plotted as a function of luminance in Figure 3. Each of the six graphs corresponds to data from a different participant. Empirical response proportions are plotted with circles and error bars (which denote standard errors[1]). The lines are model fits, to be discussed later.

Overall, there are some aspects of the response probability data that were common to several participants. Four of the 6 participants (B.G., N.C., S.E.H., and S.B.) produced response probabilities that were similar to the rules pattern in Figure 2B. Extreme stimuli (very dark or very light) were more often classified in Category A and moderate stimuli in Category B. Participants V.B. and L.T. differed from the other 4 in that their Category A response proportions were higher for moderately dark stimuli than for the darkest stimuli. Participant V.B.'s data resemble the exemplar pattern in Figure 2C. Participant L.T.'s data do not resemble either the rule or the exemplar pattern. To better gain insight on how these patterns can be interpreted in a competitive test, we fit GRT and GCM.

Model-Based Analyses of Experiment 1

Parameter Estimation

We fit models to data by minimizing the chi-square goodness-of-fit statistic with repeated application of the Simplex algorithm (Nelder & Mead, 1965). The Simplex algorithm is a local minimization routine and is susceptible to settling on a local minimum. To help find a global minimum, we constructed an iterative form of Simplex. The best fitting parameters from one cycle were perturbed and then used as the starting value for the next cycle. Cycling continued until 500 such cycles resulted in the same minimum. The amount of perturbation between cycles increased

[1] The standard errors in the figures were calculated by computing response proportions on a block-by-block basis (a block is 96 trials and consists of 12 observations at each luminance level). The variability of these response proportions across blocks was then used in calculating standard errors. This method was implemented because it is conceivable that true underlying response probabilities are not constant across different blocks or sessions. Factors that could increase the variability in response probability are variability in attention and strategy. The current method assumes that response proportions are constant within a block but vary across blocks.

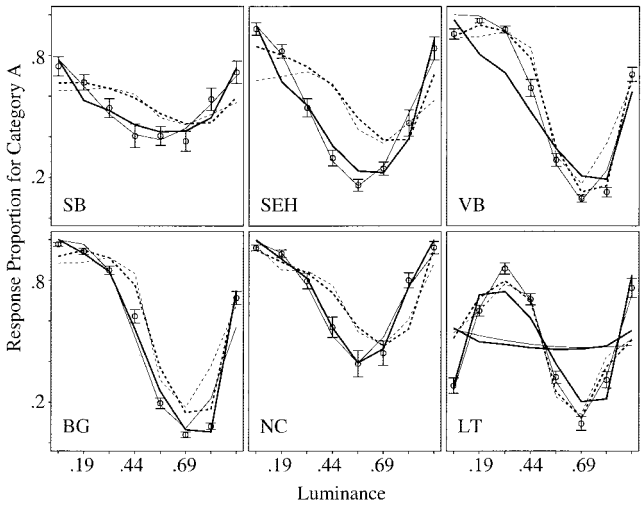

Figure 3. Category A response proportions and model fits as a function of luminance for 6 participants in Experiment 1. The circles with error bars denote participants' response proportions (error bars denote standard errors; see Footnote 1). The dashed lines denote generalized context model performance with the Absolute-Identification Scale (thick and thin lines denote performance for Gaussian and exponential gradients, respectively). The solid lines denote general recognition theory model performance (thick and thin lines denote performance for the Fixed-Variance Scale and Linear Scale, respectively). See text for a discussion of the data and predictions for Participant L.T.

each time the routine yielded the same minimum as the previous cycle.

There are some difficulties introduced by adopting a chi-square fit criterion having to do with small predicted frequencies. For several response patterns, reasonable parameters yield exceedingly small predictions, for example, .01 of a count. These small numeric values were the denominator of the chi-square statistic and led to unstable values. It is often difficult to minimize functions with such instabilities. As a compromise, the denominator of the chi-square statistic was set to the larger of 5 or the predicted frequency value.[2] Although this modification allows for more robust and stable parameter estimates, it renders the chi-square test inappropriate for formal tests. In this report, we use the chi-square fit statistic as both a descriptive index and a metric for assessing the relative fit of comparable models rather than as an index of the absolute goodness of fit of a model. We still report $p < .05$ significance values throughout for rough comparisons of fit.

Experiment 1A: Scaling Luminance

To fit both GCM and GRT, it is necessary to scale the physical domain (luminance) into a psychological domain (perceived luminance). To provide data for scaling luminance, we performed Experiment 1A, which consisted of an absolute-identification and a similarity-ratings task.

Participants. Three Northwestern University students served as participants and received $8 in compensation. Two of the 3 participants had also participated, about 6 months earlier, in Experiment 1.

Stimuli. Stimuli were the eight square patches of varying luminances used in Experiment 1.

Design and procedure. Each participant first performed the similarity-ratings task and then performed the absolute-identification task. In the similarity-ratings task, the 3 participants were given pairs of square stimuli and asked to judge the similarity of the pair using a 10-point scale. Participants observed all possible pairings five times. After completing the similarity task, participants were given an absolute-identification task in which they had to identify the eight different luminance levels by selecting a response from 1 (*darkest*) to 8 (*brightest*). After each response, participants were shown the correct response. Participants observed three blocks of 120 trials each. The last two blocks (240 trials, or 30 trials per stimulus) were retained for constructing perceived luminance scales.

Scale Construction

The data from the 3 participants in Experiment 1A were used to construct five scales of perceived luminance (presented below). Multiple scales were constructed so that we could determine whether a failure of a model was due to misspecification of scale. To foreshadow, when a model failed, it failed for all applicable scales. Hence, these failures may be attributed to structural properties of the models rather than scale misspecifications.

Similarity-Ratings Scale for GCM. In GCM, categorization decisions are determined by similarity. Data from the similarity task were converted to distance ($d = \ln[s - 1]$; Shepard, 1957, 1987) and scaled with a classic multidimensional scaling routine. Scaling was done on each of the 3 participants' data separately. For each participant, the first dimension accounted for an overwhelm-

[2] As this minimum value of the denominator is decreased, responses with probability near 0 or 1 have increasingly greater leverage in determining the fit. To see how this works, consider the case in which the predicted response probability is .5 versus the case where it is .001 (assuming no minimum denominator). There are about 300 observations per condition in the reported experiments. Therefore, the predicted frequencies are 150 and .3 for predicted probabilities of .5 and .001, respectively. Because these predicted frequencies enter in the denominator of the chi-square fit statistic, the net effect is a differential weighting of the squared difference between observed and predicted frequencies. In the case for which the predicted frequency is .3, the squared differences are weighted by a factor of 500 more than the case for which the predicted value is 150. Having a minimum value eliminates the case in which squared differences from small predicted response frequencies receive such extreme weightings. In the current example, the minimum value of 5 would be used in the chi-square denominator instead of the predicted frequency of .3. In this case, the relative weighting of the extreme response probability is only 30 times that of intermediary response probability. An alternative approach is to minimize the root-mean-square error of the response proportion predictions (see Massaro, Cohen, Campbell, & Rodriguez, 2001; Rouder & Gomez, 2001). This approach places equal weights on deviations from all response probabilities. The disadvantage of this approach is that such a method fails to take into account that extreme response proportions are associated with smaller error bars. The current technique of using a chi-square fit statistic with a minimum strives at a compromise. Extreme responses with smaller standard errors carry greater weight in determining model parameters than moderate response proportions. Because a minimum is used in the denominator, extreme response proportions do not fully determine the fit.

ing amount of variance—eigenvalues of the first dimension were more than seven times greater than those of the second. Therefore, a unidimensional scale is justified, and the coordinates on the first dimension were averaged across participants to produce the Similarity-Ratings Scale.

Absolute-Identification Scale for GCM. The absolute-identification data can also be used to construct a perceived luminance scale. To construct such a scale for GCM, we followed Nosofsky (1986) and used a variant of Luce's similarity choice model (Luce, 1959, 1963).[3]

Free-Variance Scale for GRT. To fit GRT, it is necessary to specify the mean μ_i and the variance σ_i^2 for each of the presented luminance levels. In accordance with Ashby and Lee (1991) we used the absolute-identification data to estimate these values. A GRT model of the absolute-identification task, shown in Figure 4, is quite similar to the GRT model of classification. Once again the perceived luminance line is divided into several regions, but each region is associated with a different identification response. For our applications, there are eight different means, eight different variances, and seven different bounds. The mean and variance for the darkest stimulus can be set to 0 and 1, respectively, without any loss of generality (this sets the location and scale of the perceived luminance dimension). We refer to this model as the *free-variance model* to emphasize that the variance of perceived luminance may vary across luminance. The Free-Variance Scale for μ_i was the average estimate of the means across participants; the Free-Variance Scale for σ_i^2 was proportional to the average estimate of variance. The constant of proportionality, denoted σ_0^2, serves as a free parameter.[4] The parameter σ_0 plays a similar role as the parameter λ in GCM—it linearly rescales perceived luminance.

Fixed-Variance Scale for GRT. We restricted the free-variance model to derive an alternative scale for GRT. In the restricted model, variances were not allowed to vary with luminance, and we term this the *fixed-variance model.* The Fixed-Variance Scale for μ_i was the average of the estimated means across participants; the Fixed-Variance Scale for σ_i^2 was a single constant for all luminance (this value, denoted σ_0^2, was a free parameter).

Linear Scale for GRT and GCM. Perceived luminance was a linear function of luminance. For GCM, variances σ_i^2 at each luminance level were set equal to a single free parameter, denoted σ_0^2. The slope of the linear function is inconsequential as both GCM and GRT have parameters that linearly rescale psychological distance (λ, σ_0).[5]

GCM Fits

With the provided scales, we fit GCM to the categorization data of the 6 participants in Experiment 1. Six different GCM models were obtained by crossing three applicable perceived luminance scales (Similarity-Ratings, Absolute-Identification, and Linear) with the two forms of similarity gradient (Gaussian and exponential). All six models yielded comparable results. Fits with the Absolute-Identification Scale were slightly better than with the other scales, and the corresponding predictions are shown in Figure 3 as dashed lines. The thick dashed lines are for the Gaussian gradient ($k = 2$); the thin dashed lines are for the exponential gradient ($k = 1$). The predictions from the other scales are omitted for clarity, but these predictions were similar to those displayed.

Parameter values and chi-square fit statistics for all six models are shown in Table 1. As can be seen, in all cases, GCM provides mediocre fits. Participants B.G., N.C., S.E.H., and S.B. all classify very dark stimuli in Category A more often than they do moderately dark stimuli. GCM misfits this aspect of the data.

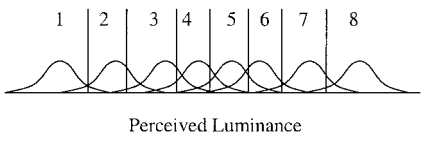

Figure 4. A general recognition theory model for absolute identification. The perceived luminance scale is partitioned into regions corresponding to the eight different stimuli.

GRT Fits

GRT was fit to the categorization data from Experiment 1 in the same manner as GCM. In total, there were three models fit, with each corresponding to a different scale (Free-Variance Scale, Fixed-Variance Scale, Linear Scale). The resultant predictions for the Fixed-Variance and Linear Scales are shown as thick and thin solid lines, respectively, in Figure 3. The predictions for the Free-Variance Scale are similar to those for the Fixed-Variance Scale and are omitted from the graph for clarity. Parameter values and chi-square values for all three GRT models are shown in Table 2.

The fits are reasonable for 4 of the 6 participants. The most noteworthy deviation occurs for participant L.T. The two solid lines that are almost horizontal and dramatically misfit the data are the three-parameter GRT predictions. For L.T., the proportion of very dark stimuli classified as Category A was lower than the proportion of moderately dark stimuli classified as Category A. This result cannot be accounted for by two-boundary GRT. However, the data suggest that L.T. may have set a third boundary for very dark stimuli. If stimuli had perceived luminance below this third boundary, then they were classified in Category B. A four-parameter (three-bound) GRT model was also fit to participant L.T.'s data, and it fit much better than the three-parameter version. The four-parameter GRT fit is shown for L.T. with the two solid

[3] The similarity choice model can be expressed in terms of the unidimensional perceived luminances:

$$P_{r_i,s_j} = \frac{B_i e^{-|d_i - d_j|}}{\sum_k B_k e^{-|d_k - d_j|}},$$

where P_{r_i,s_j} is the probability of identifying stimulus j as stimulus i, and the sum is over all available responses. The parameters d denote the perceived luminance scale values, and the parameters B denote the response biases toward particular responses. Without any loss of generality, $\sum_j B_j = 1$ (sum over available responses) and $\sum_j d_j = 0$ (sum over stimuli).

[4] The variance of perceived luminance in the absolute-identification task may not be equal to that in the categorization task. The feedback in the former is different from that in the latter—in particular, it is easier for the participants to learn the number of relative luminances of the stimuli in the absolute-identification task than in the categorization task. The free parameter σ_0^2 captures these differences.

[5] The slope was 1 unit per stimulus.

Table 1
Model Parameters and Fits for the Generalized Context Model in Experiment 1

	Exponential gradient				Gaussian gradient			
Participant	λ	γ	ϕ	χ^2 fit[a]	λ	γ	ϕ	χ^2 fit[a]
				Similarity-Ratings Scale				
S.B.	2.36	0.42	−0.15	103.46	4.03	0.39	−0.16	95.63
S.E.H.	2.58	0.88	0.00	403.11	4.11	0.87	−0.05	363.58
V.B.	2.11	2.78	−0.16	177.80	3.72	2.47	−0.13	52.68
B.G.	2.39	2.56	0.04	331.54	3.80	2.42	−0.03	184.88
N.C.	1.28	1.39	−1.16	287.59	2.80	0.79	−0.76	264.02
L.T.	5.88	0.80	0.10	254.91	16.17	0.90	0.10	231.94
				Absolute-Identification Scale				
S.B.	0.63	0.53	−0.17	99.95	0.26	0.55	−0.20	88.73
S.E.H.	0.68	1.00	−0.12	404.86	0.25	1.26	−0.32	333.46
V.B.	0.64	2.87	−0.13	211.85	0.31	3.02	−0.02	55.53
B.G.	0.65	2.73	−0.15	365.29	0.29	3.04	−0.06	175.87
N.C.	0.57	1.07	−0.73	207.41	0.39	0.69	−0.61	178.95
L.T.	1.47	1.05	0.17	220.84	1.01	1.06	0.19	186.67
				Linear Scale				
S.B.	0.54	0.46	−0.16	108.63	0.27	0.37	−0.17	105.56
S.E.H.	0.61	0.93	−0.03	428.33	0.28	0.85	−0.07	405.12
V.B.	0.57	2.75	−0.18	251.63	0.30	2.36	−0.19	127.64
B.G.	0.60	2.65	0.01	420.40	0.28	2.37	−0.10	296.30
N.C.	0.35	1.26	−1.10	317.97	0.22	0.70	−0.74	318.90
L.T.	1.53	0.83	0.12	250.46	0.86	1.01	0.17	208.90

[a] The $p < .05$ critical value of $\chi^2(5)$ is 11.07.

lines that fit the data reasonably well. For the remaining participant, V.B., GRT fit well but failed to capture the small downturn for the darkest stimuli.

Discussion

The qualitative trends from 4 of the 6 participants are highly consistent with the pattern expected from rule-based processing, and the GRT rule-based model provides a more satisfactory account than GCM for 5 of the 6 participants. Overall, the data are more consistent with a rule-based account than an exemplar-based account.

One aspect of the model-fitting endeavor is that the chi-square fit statistics for both models are quite high. Technically, both models fit poorly. However, the assumptions underlying the use of the chi-square statistic are most likely violated. One of these assumptions is that each observation is independent and identically distributed. We believe this assumption is too strong, as it is plausible that there is variability in the true response proportions from variation in attention or strategy. This variability inflates the chi-square fit statistic. We discussed a method of computing error bars in Footnote 1 that does not rely on the identically distributed assumptions—it computes variability on a block-by-block basis. If we use the error bars as a rough guide, the GRT model provides a fairly good account of the data. The GRT model predictions are within a standard error for most of the points. Conversely, the GCM predictions are outside two standard errors (an approximate 95% confidence interval) for most points.

The stimuli used in Experiment 1 were relatively confusable. It is our working conjecture that stimulus confusion may play a salient role in categorization behavior. If the stimuli are confusable, participants are not sure where to place a stimulus in the luminance space; therefore, that stimulus serves as a poor exemplar for subsequent trials. If this is the case, then participants may use rule-based categorization instead of exemplar-based categorization. In the next three experiments we manipulate stimulus confusion and examine the effects on categorization performance. To foreshadow, stimulus confusion does play a role; whereas GRT better describes performance when stimuli are confusable, GCM better describes performance when stimuli are less confusable.

Our stimulus confusion hypothesis was an outgrowth of the pilot experiments we had performed in designing Experiment 1. We found that if we presented many stimuli or placed the stimuli too close together, participants' response patterns had many Category A responses for dark stimuli and many Category B responses for light stimuli. The pattern was monotonic with Category B response proportions increasing with luminance. Initially, we considered the production of such a pattern a nuisance, as it did not fit into either of the two prototypical patterns for exemplar- and rule-based decision making (i.e., those in Figure 2). Experiment 1 was designed to avoid the production of this pattern. When participants did indeed produce this pattern, we instructed them to search for a better solution (see the *Instructions* section of Experiment 1). If stimulus confusion does play a role, then the monotonic pattern may characterize behavior when the stimuli are most confusable.

Table 2
Model Parameters and Fits for the General Recognition Theory in Experiment 1

Participant	Lower bound	Upper bound	σ_0	Additional third bound	χ^2 fit[a]
Luminance scaled by free-variance model					
S.B.	−2.34	2.54	2.53		43.1
S.E.H.	−1.72	2.71	1.48		42.9
V.B.	−0.68	4.33	1.82		265.7
B.G.	−0.57	3.96	1.15		73.5
N.C.	−0.76	1.71	1.28		66.2
L.T.	−0.16	3.25	1.19	−3.56	138.7[b]
Luminance scaled by fixed-variance model					
S.B.	−1.93	3.09	3.13		21.7
S.E.H.	−1.43	2.93	1.74		45.4
V.B.	−0.60	4.41	1.85		259.4
B.G.	−0.53	4.23	1.12		36.3
N.C.	−0.60	1.53	1.22		77.6
L.T.	−0.19	3.67	1.41	−3.40	97.9[b]
Linear Scale					
S.B.	−1.73	2.35	2.35		7.5
S.E.H.	−1.34	2.45	1.34		8.0
V.B.	−0.12	3.21	0.97		239.3
B.G.	−0.43	3.32	1.04		114.1
N.C.	−0.58	1.69	1.27		39.1
L.T.	0.00	2.93	0.94	−2.89	7.0[b]

[a] The $p < .05$ critical value of $\chi^2(5)$ is 11.07. [b] Fits to L.T. were with additional third bound.

In fact, it characterizes the simplest use of rules, a single bound on one dimension.

Our postulated role of stimulus confusion is shown in Figure 5. We suspect that in the most confusable conditions, participants opt for the simplest rule—that of a single bound. With less confusion, they may be able to place two bounds. With the least amount of confusion, participants can accurately discriminate and remember individual exemplars and use these for categorization decisions.

Experiment 2

The goal of Experiment 2 was to assess the role of stimulus confusion. The stimuli in Experiment 2 were squares of various sizes. The size determined the category assignment probability.

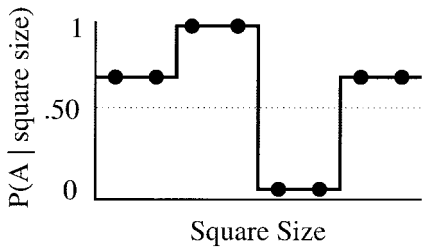

Figure 6. The category assignment probabilities for Experiment 2.

These probabilities were similar to that of Experiment 1 and are shown in Figure 6. To further motivate participants, we assigned the most extreme stimuli a probability of .7 of belonging to Category A, whereas in the prior experiments this probability was .6.

Stimulus confusion was manipulated by controlling the increment between adjacent square sizes. In the most confusable condition, participants categorized squares of eight different sizes from 70 pixels to 105 pixels. The increment from one square size to the next was 5 pixels. In the least confusable condition, participants categorized squares of eight different sizes from 30 pixels to 135 pixels with an increment of 15 pixels.

One other important difference existed between Experiments 1 and 2. In Experiment 1, we gave participants an additional instruction if, after the first session, they displayed a monotonic response pattern in which low-luminance stimuli were assigned to one category and high-luminance stimuli to the other. The additional instruction was to search for a better strategy than a monotonic pattern. In the current experiment, we did not give participants such an instruction. Participants were not discouraged from any pattern of response.

Method

Participants

Seventy-three University of Missouri students participated in Experiment 2 for course credit in an introductory psychology course.

Apparatus

The experiment was conducted on PCs with 17-in. (43.18-cm) monitors programmed to a 640 × 480 square-pixel mode. Square sizes are reported in pixels throughout. On these monitors, each pixel was 0.50 mm in length. Participants sat approximately 20 cm from the screen.

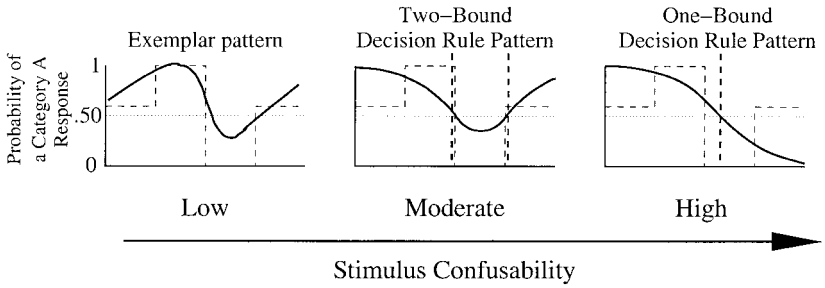

Figure 5. The relationship between stimulus confusion and categorization response patterns.

Design

The design of Experiment 2 was an 8 × 4 design. The two main factors were square size and increment between square size. Square size was analogous to luminance in the previous experiments and was manipulated in a within-participant manner. The increment between square sizes was manipulated in a between-participant manner. There were four different increments: 5, 7, 10, and 15 pixels. When the increment was 5, the square sizes were 70, 75, 80, 85, 90, 95, 100, and 105 pixels. For the other increments, the fifth largest square was always fixed at 90 pixels. For example, when the increment was 15, the square sizes were 30, 45, 60, 75, 90, 105, 120, and 135 pixels.

Procedure

Trials began with a blank display that lasted 500 ms. Then a square was presented until the participant responded and received feedback. Auditory feedback followed the response. Participants received a pleasant tone doublet (a 50-ms, 500-Hz tone followed by a 50-ms, 1000-Hz tone) for each correct response and a less pleasant buzz sound (a 100-ms, 100-Hz tone) for each wrong response. Feedback here differed from Experiment 1 in that it (a) was auditory rather than visual and (b) was presented for all trials rather than just for incorrect responses. After feedback, a rapid sequence of displays with several squares in several locations was presented. This rapid sequence did not serve as a mask as it was shown after the response and feedback. Its purpose was to reduce trial-by-trial dependencies. Because of the sequence, participants could not retain a perceptual image of the previous trial's square and, therefore, could not use it as a referent. There were 96 trials in a block and 10 blocks in a session. Participants took part in two sessions.

Participants were told not to dwell on any one trial but not to go so fast as to make careless motor errors. As in Experiment 1, participants were given a short story about probabilistic outcomes in medical tests. Unlike Experiment 1, they were not instructed about the reliability of feedback across the range of square sizes.

Results

Data Exclusion

The data from the first session for each participant were discarded as practice; only the data from the second session were used in the analysis. The first 10 trials of the second session and the first trial of each block were also discarded. All data from a participant were discarded if more than 13% of responses were outside of the 200-ms to 3-s interval. Of the 73 participants, 12 were discarded. Of these 12, 7 were discarded for excessive responses outside the 200-ms to 3-s interval, and 5 withdrew their participation by not attending the second session. Participants were excluded from the conditions in fairly equal rates. After exclusion, there were 15, 16, 14, and 16 participants in the four different increment conditions, respectively. Of the remaining 61 participants, about 1.8% of response times fell outside of the 200-ms to 3-s interval, and these trials were also discarded.

Empirical Trends

The data from the 61 participants had great variability. Figure 7 shows data from selective and representative participants. There were three observed systematic patterns. First, several participants' patterns were monotonic; they categorized small squares into Category A and large squares into Category B. The top row of Figure 7 shows data from 2 participants who exhibit the monotonic

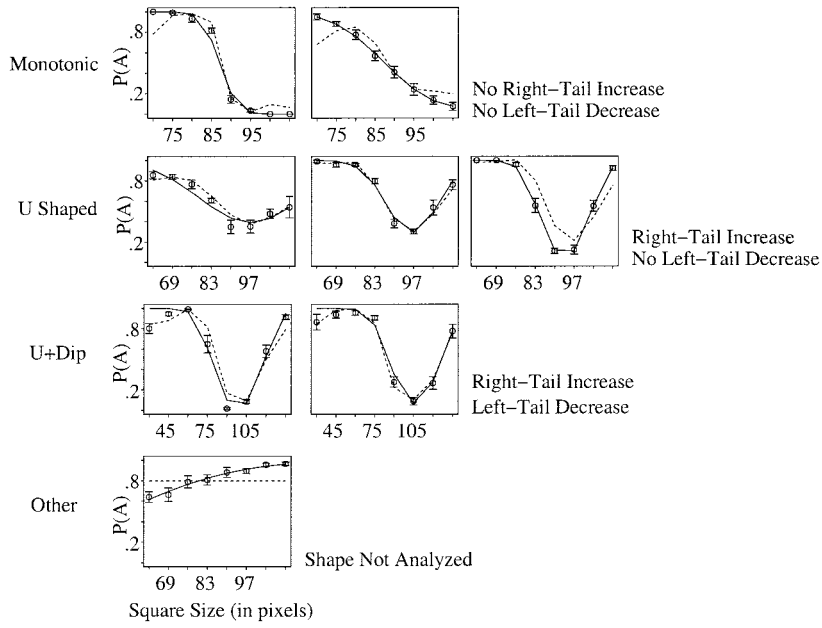

Figure 7. The types of response patterns observed in Experiment 2. Each panel shows response proportions and model predictions for selected participants. Circles with error bars denote response proportions and standard errors, respectively. The solid lines denote predictions from a general recognition theory model with the Linear Scale. The dashed lines denote predictions from a generalized context model with the Gaussian similarity gradient and the Absolute-Identification Scale.

pattern. Second, other participants' patterns were U-shaped—both very large and very small squares were classified in Category A, whereas intermediary squares were classified into Category B. The second row shows participants' data that exhibit shallow, moderate, and deep U-shaped patterns, respectively. Third, some participants produced the U-shaped pattern with a small decrease in the Category A response proportion for the smallest squares (representative data shown in the row labeled "U + Dip"). This three-part taxonomy of patterns characterized the data of 58 of the 61 participants. The other three were not well characterized by this taxonomy. An example of data that did not fit into the taxonomy is shown in the row labeled "Other." Data from these 3 participants were excluded from the analysis of empirical trends below but were retained for model-based analysis.

Although the data are diverse, there are systematic changes across square-increment condition. Two post hoc measures of the patterns were constructed to help describe these systematic changes, and these are shown in Figure 7. One measure is the degree to which response proportion increased for the largest square size. This measure is termed *right-tail increase.* Right-tail increase is constructed as a ratio. The numerator is the increase in the response proportion for the largest square (rightmost point) above the smallest response proportion. The denominator is the range of response proportions. Right-tail increase is diagnostic of whether the response pattern is monotonic with square size (corresponding to a zero right-tail increase) or U-shaped (corresponding to a high right-tail increase). The other measure is the amount that response probability dips for the smallest squares—it is termed *left-tail decrease.* Left-tail decrease is also constructed as a ratio. The numerator is the difference between the largest response proportion and that for the smallest square (leftmost point); the denominator is the range of response proportions. Left-tail decrease indicates a U + Dip pattern.

There was a general pattern across the four conditions. Monotonic patterns were more prevalent in the low square-increment conditions, whereas U-shaped ones were prevalent in the intermediate and high square-increment conditions. U + Dip patterns emerged in the higher increment condition (8 of the 16 participants in the largest increment condition had some degree of left-tail decrease). Table 3 shows the mean right-tail increase and left-tail decrease as a function of condition. Although these measures are continuous in nature, there are many participants who scored near the bounds of one and zero. In this case, nonparametric statistics are useful in assessing trends. Each participant was ranked on both measures. Table 3 also shows the mean ranks for each increment condition (each mean rank is from 58 participants). Both right-tail increase and left-tail decrease increase with increment. The direction and monotonicity of these trends support the hypothesis that stimulus confusion affects categorization behavior with a progression from simple rules to more complicated rules to exemplar storage with increased stimulus confusion. To assess the statistical significance of these trends, we performed a planned comparison in which ranks in the two low-increment conditions were compared against ranks in the high-increment conditions with a Mann–Whitney U test (Hayes, 1994). The changes were significant (right-tail increase: $U = 242.5$, $p \leq .05$; left-tail decrease: $U = 296$, $p \leq .05$). To further test the working hypothesis about the role of stimulus confusion, we conducted model-based analyses.

Table 3
Effect of Increment Conditions in Experiment 2

	Increment			
Measure	5	7	10	15
	Means			
Right-tail increase	.31	.45	.67	.65
Left-tail decrease	.01	.01	.03	.05
	Mean ranks			
Right-tail increase	20.0	26.5	35.5	35.8
Left-tail decrease	24.8	25.6	29.8	37.1

Model-Based Analyses of Experiment 2

To perform model analyses, it is necessary to scale the square sizes into a psychological space of perceived size. Fifteen University of Missouri undergraduates performed an absolute-identification task similar to that of Experiment 1A for this purpose. With these absolute-identification data, Absolute-Identification, Free-Variance, and Fixed-Variance Scales were constructed as previously discussed.

Four GCM models were fit to the categorization data of Experiment 2. These models were obtained by crossing the two perceived-size scales (Absolute-Identification Scale and Linear Scale) with the similarity gradients (exponential and Gaussian). Although all four model fits were fairly similar, GCM with the Gaussian similarity gradient and Absolute-Identification Scale fit the best for three of the four increment conditions. To keep the graphs less cluttered, only these predictions are shown (Figure 7, dashed lines). GCM models not displayed yielded qualitatively similar predictions and misfit aspects of the data in the same manner as the displayed predictions. Three GRT models were fit (Fixed-Variance Scale, Free-Variance Scale, and Linear Scale). The fit with the Linear Scale was best in each of the four increment conditions. The corresponding predictions are shown (Figure 7, solid lines). GRT fits with Fixed-Variance and Free-Variance Scales were similar to that with the Linear Scale; misfits occurred at the same points and in the same direction.

The results of the model-based analyses are fairly straightforward. If a participant displayed a monotonic pattern with smaller squares placed in Category A and larger squares placed in Category B, then only the GRT model provided a satisfactory fit. GRT in these cases reduced to a one-bound model as the estimated upper bounds were much greater than the largest square size. The fits to the U-shaped pattern were more complex. If the U-shaped pattern was relatively shallow without extreme response proportions, then GRT tended to fit better than GCM. However, if the U-shaped pattern was relatively deep with extreme response proportions, then both GCM and GRT fit well. If there was some degree of left-tail decrease, then GCM outperformed GRT.

Overall, GRT fit better in the small increment, high stimulus-confusion conditions, as participants exhibit monotonic and U-shaped patterns. But in the lowest stimulus-confusion condition, a substantial number of participants demonstrate left-tail decrease, and these participants' data are best fit with GCM. Figure 8

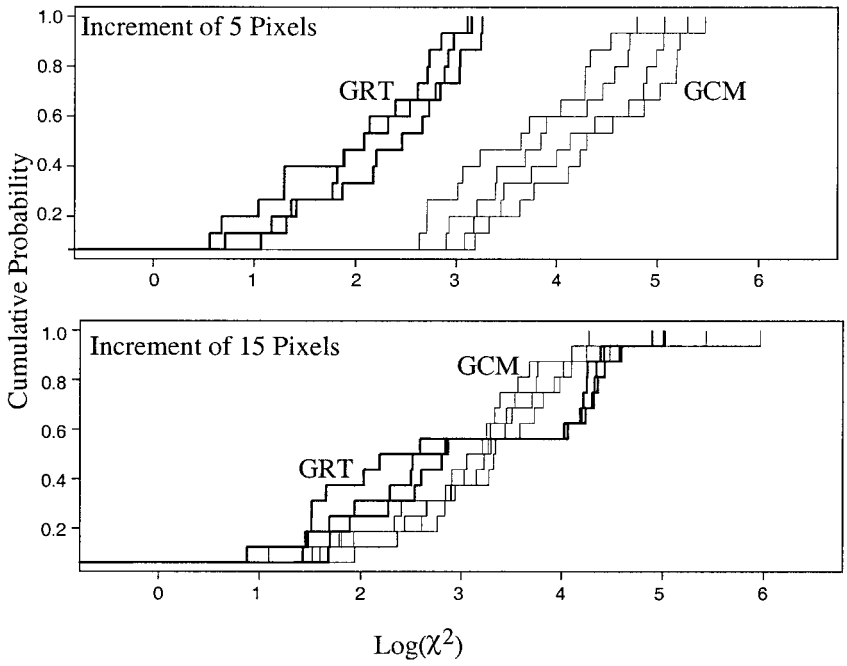

Figure 8. Cumulative distribution functions of the natural logarithm of chi-square fit statistics for square-size increment conditions of 5 and 15 in Experiment 2. The cumulative probability is obtained by ordering participants' chi-square fit statistics. The three thick lines denote general recognition theory (GRT) model fits; the four thin lines denote generalized context model (GCM) fits.

graphically depicts these trends. It shows the distributions of the chi-square fit statistics as empirical cumulative distribution functions. The chi-square values of each participant were ordered from smallest to largest; the logarithms of these values are plotted on the *x*-axis. The *y*-axis is the proportion of fits less than or equal to a certain value. The cumulative distribution functions of chi-square statistics from well-fitting models tend to dominate those from poor-fitting models. In the top panel, for a square size increment of 5, the distribution of chi-square values from the three GRT models is shown as solid lines. These dominate the four dashed lines from the four GCM models. This dominance indicates a better fit for GRT than for GCM. But for an increment of 15, none of the three GRT models dominates the four GCM models. For this condition, 10 participants are better fit by GCM, whereas 6 are better fit by GRT. Those participants better fit with GCM tend to have substantial left-tail decrease.

Discussion

Participants exhibited a wide variety of categorization behavior. Even so, the empirical trends and model-based analysis tell a consistent story: Stimulus confusion affects categorization performance. When the stimuli are most confusable, participants tend to produce a monotonic pattern that is better fit by GRT. These good-fitting GRT models have a single bound. As stimulus confusion is lessened, more participants tend to produce more U-shaped patterns. These U-shaped patterns also tend to be better fit with GRT. In the least confusable conditions, there is a left-tail decrease for some participants, and these patterns are better fit by GCM than by GRT. The main caveat is that there is much individual variability. A set of stimuli that are confusable for one participant may not be for another. Hence, these two participants may use different strategies. Our finding that different individuals are better fit by different models is not too surprising and is in accordance with the results of Nosofsky et al. (1994) and Thomas (1998).

We suspect that stimuli in the least confusable condition were still somewhat confusable. There is a well-known processing limitation in absolute identification with unidimensional stimuli. Pollack (1952) found that participants were only 80% accurate in identifying 14 tones between 500 Hz and 1000 Hz. Pollack then increased the range (14 tones were placed between 500 Hz and 5000 Hz) but still obtained only 80% accuracy. This result, along with several similar ones in different unidimensional domains, suggests that there is a mnemonic limitation in the ability to identify unidimensional stimuli that cannot be abrogated by increasing the physical separation between stimuli (see Miller, 1956; Shiffrin & Nosofsky, 1994). Rouder (2001) found that absolute identification with line lengths is fairly limited with six lines. These studies, taken together, indicate that there may be some mnemonic confusion for a set of eight square sizes, even for arbitrarily large increments. Stimulus confusion reflects contributions from perceptual variability and mnemonic limits. In short, stimulus confusion is tantamount to the ability to absolutely identify stimuli. If stimuli cannot be identified in an absolute sense, then there is at least some degree of stimulus confusion. To further explore the role of stimulus confusion, we made the stimuli less

confusable in Experiment 3 by using only four well-spaced squares. This stimulus set yields reliable absolute identification and greatly reduced stimulus confusion.

Experiment 3

Experiment 3 was designed to assess categorization performance when stimulus confusion is greatly reduced. Reduction in stimulus confusion was obtained by using four well-spaced squares. The category assignment probabilities for the squares are shown in Figure 9. Although this assignment pattern differs from the previous experiment, it retains the important property of providing for differential predictions. Exemplar-based models predict that there should be more Category A responses to Stimulus II than to Stimulus I because there are more Category A exemplars for Stimulus II. Rule-based models predict a bound between Stimulus III and Stimulus IV. Stimuli smaller than the bound (Stimuli I, II, and III) should be classified into Category A on a majority of trials, whereas Stimulus IV, the sole stimulus larger than the bound, should be classified into Category B on a majority of the trials. Stimulus II is closer to the bound than is Stimulus I. Accordingly, there should be more Category A responses for Stimulus I than for Stimulus II.

Method

Participants

Six University of Missouri students participated in the experiment. They were compensated $7.50 per experimental session. They each participated in two sessions.

Stimuli and Design

The size of the square was manipulated (within block) through four levels: 15 pixels, 35 pixels, 60 pixels, and 90 pixels. The category assignment probabilities for Stimuli I, II, II, and IV were .583, .767, .933, and .067, respectively.

Procedure

The procedure was nearly identical to that of Experiment 2. Participants observed twice as many trials per stimulus as in Experiment 2, as there were half as many stimuli.

Results

Analysis was on data from the second session. The first 10 responses and the first response of each block were discarded. About .3% of the remaining responses fell outside of the 200-ms to 3-s interval. These responses were also discarded. The resulting response proportions for the four size conditions are shown in Figure 10. All 6 participants classified Stimulus II in Category A more often than they did Stimulus I, hence qualitatively violating the decision-bound model. Consider as a null hypothesis that participants were as likely to classify Stimulus I into Category A as they were to classify Stimulus II. The probability that all 6 participants would classify Stimulus I into Category A more often than they did Stimulus II under this null hypothesis is less than .016. Hence, we can generalize the result even though it was obtained with a relatively small sample size.

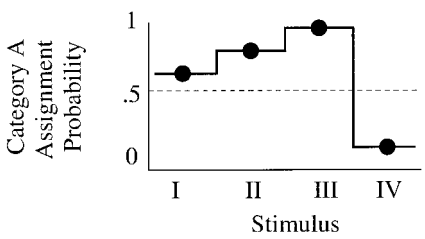

Figure 9. Category assignment probabilities for Experiment 3.

Model-Based Analyses of Experiment 3

Scaling Luminance

To fit both GCM and GRT, it is necessary to scale the stimuli into a psychological space. Three undergraduates served as participants in exchange for course credit. They performed similarity-ratings and absolute-identification tasks similar to those in Experiment 1A. A Similarity-Ratings Scale was constructed with the same method as in Experiment 1A. We were not successful in scaling the stimuli with the absolute-identification task and could not produce Absolute-Identification, Free-Variance, or Fixed-Variance Scales. The reason for this failure is that participants performed near ceiling (.98) in the absolute-identification task. The lack of confusions made distance estimates unreliable. Of course, this is expected, as the stimuli were so few in number and distinct. The use of a single task-based measure is not problematic, as the model-based conclusions in the previous experiments have been robust to differences in the scales.

In addition to the Similarity-Ratings Scale, two other scales were constructed. One was the Linear Scale, in which perceived size was linear with veridical size. The other scale was linear with stimulus's ordinal position (i.e., 1, 2, 3, 4) and is termed the Ordinal-Position Scale. In the previous experiments, the Linear Scale and the Ordinal-Position Scale were the same, as square size (luminance) was linearly related to ordinal position. In this experiment, the squares were not sized in a linear fashion; instead, the difference between consecutive squares increased with square size.

GRT Fits

The optimal solution from the GRT perspective is to use a single bound (it should be set between Stimuli III and IV). We used this solution as guidance and implemented a one-bound, two-parameter GRT model. Two such models were fit: one with the Linear Scale and the other with the Ordinal-Position Scale. The model fared poorly with both scales (chi-square goodness-of-fit values are shown in Table 4). The GRT predictions (with the Linear Scale) are shown as solid lines in Figure 10. The predictions for the Ordinal-Position Scale are similar in that they misfit the same points and in the same directions.

GCM Fits

We used a two-parameter version of GCM. This was to equate the number of parameters in both GCM and GRT. Model comparisons are easier to make if both models have the same number of parameters. A two-parameter GCM model was implemented by

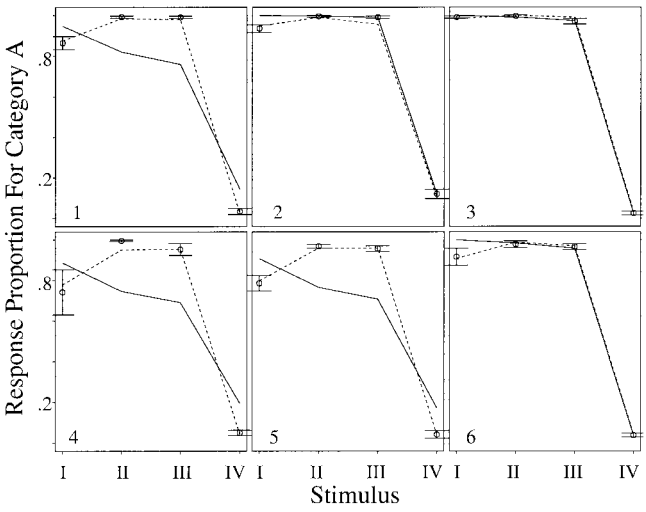

Figure 10. Category A response proportions and model fits as a function of square size for 6 participants in Experiment 3. The circles with error bars denote participants' response proportions (error bars denote standard errors; see Footnote 1). The solid lines denote predictions from a general recognition theory model with the Linear Scale. The dashed lines denote predictions from a generalized context model with the Gaussian similarity gradient and the Linear Scale.

setting response bias (ϕ) to zero. Six GCM models were fit; these were obtained by crossing the three scales (Similarity-Ratings, Linear, and Ordinal-Position) with the two similarity gradients (exponential and Gaussian). It mattered little which GCM models we fit. GCM predictions (Gaussian similarity and Linear Scale) are shown as dashed lines in Figure 10, and the chi-square goodness-of-fit values are shown in Table 4. As can be seen, GCM does a better job of fitting the data than GRT. This relation holds regardless of the scale or similarity gradient.

Discussion

The results and model analysis show that Experiment 3 is more consistent with GCM than GRT. The most salient difference between this experiment and the previous ones is the reduced number of stimuli and subsequent reduced stimulus confusion. As is consistent with the previous results, this reduction in confusion allows for the placement of exemplars and a shift to exemplar-based categorization.

It may be possible to account for the data with GRT in a post hoc fashion. Perhaps participants adopted a second decision bound *below* the smallest square size. If this is the case, due to perceptual variability, some of the presentations from the smallest square stimulus may fall below this bound and elicit a Category B response. This two-bound GRT model can account for the left-tail decrease exhibited by the participants. But such an explanation seems implausible as it postulates a decision bound in a region of the space in which no stimuli are presented. The following experiment was designed to further test the stimulus confusion hypothesis. It was also designed in a manner to preclude a bound placement outside the stimulus range.

Experiment 4

The results of the preceding experiments provide evidence for the hypothesis that stimulus confusion plays a role in determining the mode of categorization. The goal of Experiment 4 is to provide further evidence for the working hypothesis, and stimulus confusion was again manipulated. In contrast to Experiment 2, the main manipulation of stimulus confusion was the addition of a line with tick marks under the to-be-classified square. With this line, an astute observer could determine the size of the square without perceptual variability and presumably without much stimulus confusion. To draw comparisons, there was a control condition in which the tick-marked line was not provided.

Participants categorized squares of six different sizes into two categories. The category assignment probabilities are shown in Figure 11. In this paradigm, the appropriate strategy in the decision-bound approach is to place a single bound. If the perceived size of a square is smaller than the bound, the square should be classified into Category A; otherwise it should be classified into Category B. The predictions for the GCM model are qualitatively more flexible; it can account for the monotonic pattern. But GCM can account for a pattern GRT cannot: the one with more moderate classification probabilities for the second and fifth stimuli than for the third and fourth stimuli. One of the advantages of this design is that the extreme stimuli also have extreme categorization probabilities. Hence, it is unreasonable to think that participants

Table 4
Chi-Square Goodness-of-Fit Values for Experiment 3

			Participant					
Model	Gradient	Scale	1	2	3	4	5	6
GCM	Gaussian	Similarity	3.3	8.5	5.7	12.0	0.8	1.5
GCM	Gaussian	Linear/Ordinal Position	2.9	8.4	5.7	11.4	0.6	1.7
GCM	Gaussian	Linear/Square Size	13.0	21.1	3.4	33.9	18.1	5.7
GCM	Exponential	Similarity	3.8	5.0	8.4	15.8	3.1	2.3
GCM	Exponential	Linear/Ordinal Position	3.6	4.9	8.4	15.4	2.9	2.4
GCM	Exponential	Linear/Square Size	12.7	8.3	7.7	24.9	8.0	1.3
GRT		Linear/Ordinal Position	321.3	46.2	1.0	364.0	327.1	86.8
GRT		Linear/Square Size	171.6	46.0	1.1	232.8	190.5	82.1

Note. GCM = generalized context model; GRT = general recognition theory.

adopted bounds outside of the range of presented square sizes. According to the stimulus-confusion hypothesis, the data will be differentially better accounted with GCM than GRT in the experimental condition (with the tick-marked line placed under stimuli) than in the control condition.

Method

Participants

Twenty-four University of Missouri undergraduate students served as participants for course credit in an introductory psychology course. Fourteen served in the experimental condition, 9 served in the control condition, and 1 was eliminated because the resulting data showed no variation across different square sizes.

Design

The design was a 6 × 2 mixed-factorial design. Square size was manipulated within participants. The six square sizes ranged from 10 pixels to 160 pixels in increments of 30 pixels. There were two levels of confusion: experimental condition, with the tick-marked line, and control condition, without the tick-marked line. Confusion was manipulated between participants.

Procedure

The procedure was identical to that in Experiments 2 and 3 with the following exception. In the experimental condition, the line with the tick mark appeared 10 pixels below the bottom edge of the to-be-classified square. The line was centered under the square. Participants in the experimental condition were instructed to use the line in making their categorization decisions. There were equal numbers of all six stimuli. As before, sessions consisted of 10 blocks of 96 trials each.

Results

Data exclusion followed previous guidelines. The first 50 trials were discarded as practice trials, as was the first trial of each block. Two participants were discarded for excessive responses outside of the 200-ms to 3-s interval. For the remaining participants, 2.0% of the response times fell outside of the interval, and these trials were discarded. The averaged categorization proportions are shown in Figure 12. In contrast with the previous three studies, data from individual participants were sufficiently similar to warrant averaging. In the control condition, participants' categorization proportions followed a monotonic pattern even though the assigned categorization proportions were nonmonotonic. By contrast, in the experimental condition, participants' categorization proportions reflected the nonmonotonicities in the assigned categorization probabilities. Analysis of individuals (Figures 13 and 14) shows

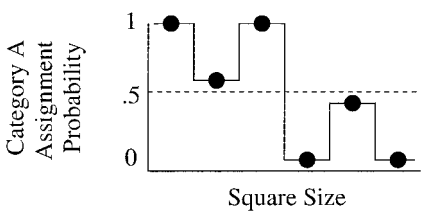

Figure 11. Category assignment probabilities for Experiment 4.

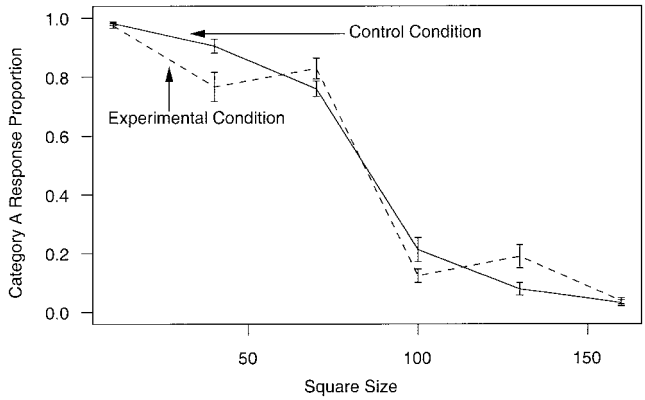

Figure 12. Results from Experiment 4. Data are means over all participants. Error bars denote standard errors and were calculated across participants.

that all 9 participants in the control condition exhibited strict monotonicity in response proportion as a function of size, whereas 10 of the 12 participants in the experimental condition exhibited at least one violation of monotonicity.

Scale Construction and Model-Based Analyses

The previous experiments indicate that the Linear Scale provides for a fair and representative comparison of the models with these stimuli. It was the only scale used in fitting the data from Experiment 4. As in Experiment 3, we fit a two-parameter version of each categorization model. For GRT, there was a single bound, and for GCM, bias was set to zero.[6] We fit two GCM models (one for each gradient) and one GRT model. Predictions are shown in Figures 13 and 14. The displayed GCM fit (dashed lines) is for the Gaussian similarity gradient; predictions from the exponential similarity gradient are nearly identical and are omitted for clarity.

The distributions of the chi-square fit statistics across participants are plotted as cumulative distribution functions in Figure 15. Overall, GCM fits the data better than GRT in both conditions. The difference in fits is much larger in the experimental condition than in the control condition. Only GCM can account for the nonmonotonic pattern that resulted from adding the tick-marked line.

It was expected that GCM would fit better than GRT in the experimental condition as there was little stimulus confusion, but GCM also fit better in the control condition. This is not surprising, as the stimuli are relatively low in number and distinct. In support of the hypothesis, GCM differentially fit better when stimulus confusion was reduced by the addition of the tick-marked line.

As noted previously, GCM is more flexible than GRT for this design, and flexibility has become a topical issue in model selection (e.g., Kass & Raftery, 1995; Myung & Pitt, 1997; Pitt, Myung,

[6] We encountered a new difficulty in fitting GCM in this experiment. The model fit well, but for some participants it yielded a very high estimate for parameter γ. This in itself is not problematic, as parameter γ denotes the degree of stochasticity in the response. High values indicate that participants are very consistent in their answers. To speed convergence, we capped the estimate of γ to be less than or equal to a value of 500, which is exceptionally high.

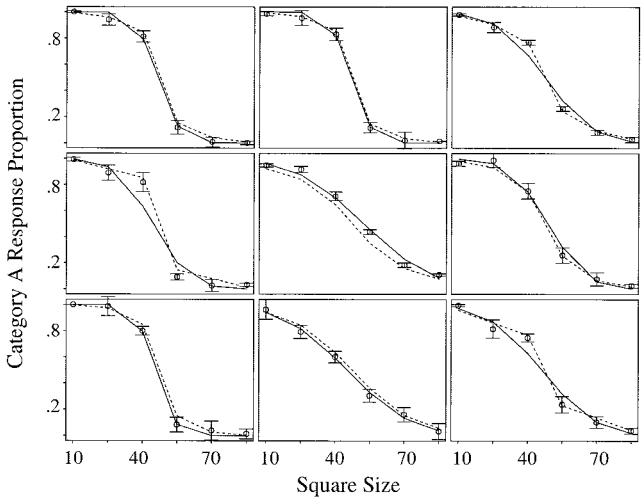

Figure 13. Response proportions and model predictions for the 9 participants in the control condition of Experiment 4. Circles with error bars denote response proportions and standard errors (see Footnote 1), respectively. The solid lines denote predictions from a general recognition theory model with the Linear Scale. The dashed lines denote predictions from a generalized context model with the Gaussian similarity gradient and the Linear Scale.

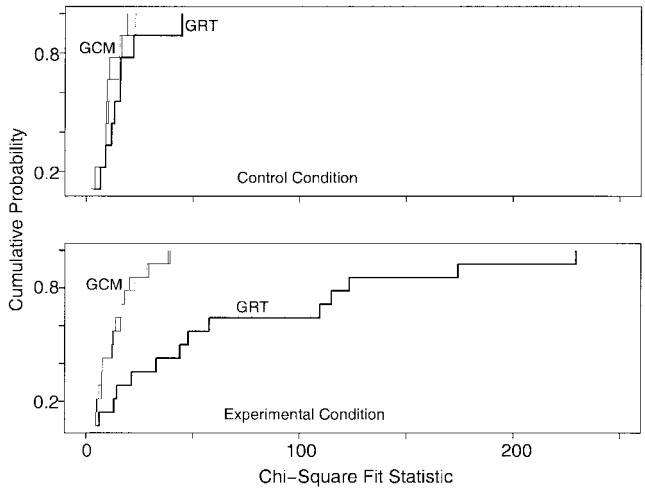

Figure 15. Cumulative distribution functions of chi-square fit statistics for control and experimental conditions in Experiment 4. The cumulative probability is obtained by ordering participants' chi-square fit statistics. The thick solid lines denote the values from the general recognition theory (GRT) model with the Linear Scale. The thin lines denote the values from the exponential and Gaussian gradient generalized context model (GCM) (both with the Linear Scale).

& Zhang, 2003). Although this difference in flexibility makes model selection complicated in cases in which both GRT and GCM qualitatively account for the data, it is of little concern for cases in which one model clearly fails. Therefore, whereas the small advantage of GCM over GRT in the control condition may be accounted for in terms of increased flexibility, the large advantage of GCM over GRT in the experimental condition may not.

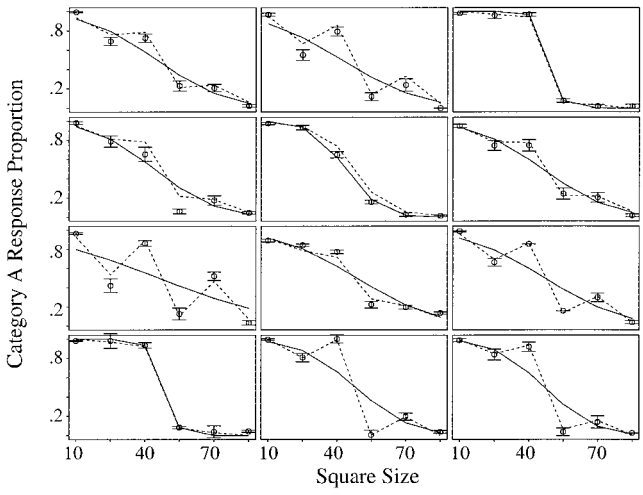

Figure 14. Response proportions and model predictions for the 12 participants in the experimental condition of Experiment 4. Circles with error bars denote response proportions and standard errors (see Footnote 1), respectively. Solid and dashed lines denote general recognition theory (GRT) and generalized context model (GCM) predictions, respectively. The solid lines denote predictions from a GRT model with the Linear Scale. The dashed lines denote predictions from a GCM model with the Gaussian similarity gradient and the Linear Scale.

General Discussion

We have presented a paradigm for contrasting rule- and exemplar-based theories of categorization behavior. The paradigm is based on probabilistic assignment—and the assignment is complex and nonlinear. Importantly, categorization theories yield differential predictions in this paradigm. The results from the four experiments, taken together, yield consistent findings. First, for any given experiment, there is a large degree of individual variability. Often, some participants produce data that are better fit by GCM, and others produce data that are better fit by GRT. This degree of participant-level variability is present in similar studies (e.g., Nosofsky et al., 1994; Thomas, 1998). Second, and more important, the degree to which GCM and GRT held varied systematically with stimulus confusion. GRT better describes categorization with confusable stimuli; GCM better describes categorization with distinct stimuli. The rationale offered is that when the stimuli are too confusable, placement of exemplars may not be sufficiently reliable for use on subsequent categorization trials. In Experiment 1, there was a moderate degree of stimulus confusion, mostly due to the use of eight, unidimensional stimuli. GRT fit better than GCM. In Experiment 2, the distance between stimuli was manipulated. Although the data were variable, the evidence suggests that rules were used in the more confusable conditions but that rules and exemplars were used in the least confusable condition. We speculate that stimuli in the least confusable condition were still confusable—owing to the nature of processing limits with unidimensional stimuli. In Experiment 3, stimulus confusion was reduced by using only four well-spaced stimuli, and as predicted, GCM better described categorization performance. Finally,

in Experiment 4, stimulus confusion was reduced by presenting participants a tick-marked line such that they could "measure" the stimuli. Although GCM fit better than GRT in both conditions, it differentially fit better in the reduced-confusion condition.

There are a number of models derived from GCM and GRT. Kruschke's (1992) attention learning covering map (ALCOVE) is a neural network instantiation of GCM that models trial-by-trial learning. As participants reach asymptotic performance, GCM and ALCOVE yield equivalent predictions. ALCOVE is challenged in the conditions in which GCM fails, notably when there is a moderate to large degree of stimulus confusion. Another model derived from GCM is Nosofsky and Palmeri's exemplar-based random walk (EBRW; Nosofsky & Palmeri, 1997). EBRW makes qualitatively similar accuracy predictions to GCM in the limit of asymptotic performance and is challenged by results GCM fails to predict. GRT has generalizations, too. Ashby (2000) recently proposed a stochastic version of GRT that accounts for response choice and response time. But this model is also challenged by the data from the low stimulus-confusion conditions that follow exemplar-like patterns.

Dual-Process/Dual-System Categorization Models

Recently, a number of researchers have advanced dual-process or dual-system categorization models, models that posit two distinct processing modes for categorization. Nosofsky and colleagues' rule-plus-exception model (RULEX) is an example of a categorization model that has both rule- and exemplar-based mechanisms (Nosofsky et al., 1994; Nosofsky & Palmeri, 1998). According to RULEX, people first use simple, one-bound, dimensional rules to solve categorization problems. If this strategy fails (i.e., if the participants believe they are making an unacceptable proportion of errors), then participants adopt more complex conjunctions of simple rules. If this strategy fails, participants may supplement the rules by memorizing a few exceptions in addition to the rules. RULEX can certainly account for the rulelike behavior observed in the experiments. But RULEX can also store exemplars and does so in a region where rules produce too many errors. For example, it is within the spirit of RULEX to posit exemplars for the extreme stimuli in Experiments 1 and 2—those with intermediate category assignment probabilities. In this manner, the model can qualitatively predict the exemplar-based data pattern. The only difficulty with RULEX is that there are no mechanisms to explain the effects of stimulus confusion. One extension would have the choice of strategy (whether to use simple rules, complex rules, or complex rules and exemplars) be dependent on stimulus confusion. With very confusable stimuli, participants expect and tolerate more errors. Therefore, they may be content with only single-criterion rules. With moderately confusable stimuli, the tolerance for error decreases and the rules become more complex. With the less confusable stimuli, tolerance of errors is at its lowest, and the rules are supplemented with exceptions.

Erickson and Kruschke's (1998) ATRIUM is another dual-process model of categorization. According to ATRIUM, participants use a rule-based module and an exemplar-based module in parallel. On each trial, both modules compute category activations. A gate then determines the net influence of each module on the categorization decision. The gate is complex and is differentially set depending on the stimulus. For stimuli that are correctly classified by a simple dimensional rule, the gate favors the rule-based module; for stimuli that are correctly classified by an exemplar, the gate favors the exemplar-based module. ATRIUM incorporates learning mechanisms for training the gate and the modules. ATRIUM can theoretically account for all three of the patterns presented in Figure 5. The single-bound solution (left panel) can be obtained by biasing the gate toward the rule-based mechanism. The exemplar-based solution (right panel) can be obtained by biasing the gate toward the exemplar-based mechanism. The two-bound solution (center panel) can be obtained by using the simple rule represented by the left bound and Category A exemplars for large square sizes (high luminance). The model is richly parameterized, and there has yet to be a discussion of how specific learning parameters influence the emergence of predictive patterns. It is unclear whether the model could predict the observed data.

ATRIUM, like RULEX, does not explicitly account for changes in classification data as a function of stimulus confusion. Erickson and Kruschke (2002) provide an informal discussion of when rules are induced in ATRIUM. They state that rules are more likely to be induced when the categorization problem is easier. Their argument goes as follows: Learning an applicable rule is a beneficial strategy because if the rule is correct, then the participant has learned the correct categorization strategy for several stimuli at once. Learning exemplars, while guaranteed to provide good categorization decisions for repeated stimuli, takes more time, effort, and resources. In ATRIUM, learning is competitive; learning rules reduce the ability to learn with exemplars and vice versa. Erickson and Kruschke argue that in the face of this trade-off, participants should adopt rules for easy tasks (presumably those without stimulus confusion) while relying on exemplars for difficult tasks (presumably those with stimulus confusion).

Although we agree that learning both rules and exemplars will happen differentially across different conditions, our account differs from that of Erickson and Kruschke (2002). We find that exemplars are used for easy tasks whereas rules are used for more difficult ones. Our explanation is in accordance with Erickson and Kruschke's observation that learning rules does simplify categorization problems greatly. We suspect that because rules do simplify problems, they will be invoked when simplification is needed most, that is, when the task is difficult (with stimulus confusion). In cases in which stimuli are distinct, then, eventually exemplars may be reliably placed and used.

Ashby and colleagues (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Ell, 2001, 2002; Waldron & Ashby, 2001) have also been persuasive in proposing a two-system categorization model. One system, the explicit system, makes categorization decisions based on simple, easy-to-verbalize rules. An example of an easy-to-verbalize rule comes from Maddox and Ashby's (1993) categorization problem for rectangles (see Figure 1C of this article). This problem can be solved with the verbal rule of classifying rectangles with heights greater than widths into Category A and those with widths greater than heights into Category B. The explicit system corresponds well with theories that participants are forming and testing different explicit hypotheses to solve the categorization task. Participants consciously maintain these rules in the explicit system; its locus is assumed to be in the frontal lobes. The second categorization system is the implicit system. Participants do not have conscious access to this system;

its locus is in the basal ganglia. The implicit system works through slower, Hebbian-like, associative learning. It can learn solutions that need not be stated verbally. For example, it is impossible to verbalize a rule that captures a reasonable decision bound in McKinley and Nosofsky's (1995) experiment depicted in Figure 1E. This problem is solved with the implicit system and not the verbal explicit system. Ashby et al. (1998) have implemented this two-system theory as both a neuropsychological model and a formal mathematical model. Both implementations are termed competition between verbal and implicit systems (COVIS).

COVIS receives support from several domains. First, it accounts for dissociations in behavioral effects. For example, the time to learn a categorization problem depends on whether the rule can be verbalized (Maddox & Ashby, 1993). It also accounts for the differential effect of Stroop-like interference (Waldron & Ashby, 2001). This interference occurs for simple categorization problems with verbalizable solutions but not for more complex ones with nonverbalizable solutions. A second line of support comes from dissociations among clinical populations. Patients with basal ganglia disease, frontal lobe damage, and temporal lobe damage show different patterns of categorization impairment (Ashby & Ell, 2001, 2002; Knowlton, Mangels, & Squire, 1996; Knowlton & Squire, 1993; Knowlton, Squire et al., 1996; Maddox & Filoteo, 2001). Finally, neuroimaging studies have shown differential patterns in brain activation depending on the particular categorization task (Patalano, Smith, Jonides, & Koeppe, 2002; Poldrack, Prabhakaran, Seger, & Gabrielli, 1999; Reber, Stark, & Squire, 1998). Each of these lines of support, however, has been critiqued in some part on logical grounds (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; Zaki & Nosofsky, 2002).

The question of interest is whether the exemplar pattern of results in Experiments 3 and 4 challenges COVIS, the formal implementation of Ashby's dual-system theory. Technically, it does. Both systems are rule based, and neither can account for the observed behavior that is inconsistent with GRT. But Ashby and Ell (2001, 2002) present an informal alternative—they believe it is possible to consciously remember a few stimulus–category pairings. Of course, remembering a stimulus–category pairing, whether conscious or not, is an exemplar process. In essence, then, Ashby and Ell describe a three-system account: an explicit rule system, an implicit rule system, and an exemplar system for a small number of stimulus–category pairings. For unidimensional stimuli, our proposal about the role of exemplars is fairly concordant with that of Ashby and Ell's alternative. If there are a few distinct stimuli, categorization will be mediated by exemplars. Alternatively, if there are several confusable stimuli, categorization is mediated by one or both of the rule-based systems. In the next section, we discuss the case of complex multidimensional stimuli. For these cases, our approach differs from that of Ashby and Ell.

Stimulus Dimensionality

Our results were obtained in a unidimensional paradigm. Although this paradigm is simple and informative, the complexity of the results makes generalization complicated. In this article, we used the term *stimulus confusion* as a proxy for absolute identification. It is well known that identification with unidimensional and low-dimensional stimuli is limited, even after extensive practice (Hartman, 1954; Miller, 1956; Rouder, Morey, Cowan, & Pfaltz, in press; Shiffrin & Nosofsky, 1994). For example, people can identify only seven or so tones and line lengths. According to our results, exemplar-based results hold only if there are less than a handful of unidimensional stimuli. If there are more than seven, then participants may not be able to place exemplars and may only use rules. The number of stimuli people can recognize in a domain does indeed increase with additional dimensions. There are a variety of estimates of how many two-dimensional stimuli people can identify. Although these estimates vary depending on procedure and stimuli, the limit of identification is less than 25 stimuli (Eriksen & Hake, 1955; Lockhead, 1970). This limit is greater than that with unidimensional stimuli, but it is still a severe limit. We hypothesize that exemplar-based performance will be seen below the limit, whereas rule-based performance will be seen above it. Accordingly, rule-based performance will be observed for categorization problems with stimulus confusion—cases in which there are many low-dimensional stimuli such as simple geometric shapes, colors, or patches of luminance.

The limitations in identification can be contrasted to domains in which stimuli are complex. People can identify as many as 100,000 different words (Landauer, 1986) and probably just as many objects. It is helpful to dichotomize domains into those that are "simple," such as line lengths, tones, colors, and other simple geometric objects, and those that are "complex," such as faces, words, and objects. With appropriate practice, absolute identification is not limited for complex stimuli under normal viewing conditions. For example, there is no magic number that describes the maximum number of words one can learn. Participants learn to minimize stimulus confusion for complex stimuli (Gibson & Gibson, 1955), whereas they do not for unidimensional stimuli. We speculate that for distinct, well-learned complex stimuli, participants can use exemplars—even when there are tens, hundreds, or thousands of stimuli. In this sense, our proposal differs from that of Ashby and Ell (2001, 2002) who believe that exemplars are used when there are only a few stimulus–category mappings. It is hoped that the paradigms presented here may be adapted for use with complex stimuli, for example, with faces or words. If so, the generality of the role of rules and exemplars in categorization may be further elucidated.

References

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. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

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., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,* 204–210.

Ashby, F. G., & Ell, S. W. (2002). Single versus multiple systems of learning and memory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 655–691). New York: Wiley.

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., & Lee, W. W. (1991). Predicting similarity and categoriza-

tion from identification. *Journal of Experimental Psychology: General, 120,* 150–172.

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., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95,* 124–150.

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

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

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review, 9,* 160–168.

Eriksen, C. W., & Hake, H. W. (1955). Multidimensional stimulus differences and accuracy of discrimination. *Journal of Experimental Psychology, 50,* 153–160.

Espinoza-Varas, B., & Watson, C. (1994). Effects of decision criterion on latencies of binary decisions. *Perception & Psychophysics, 55,* 190–203.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32–41.

Hartman, E. B. (1954). The influence of practice and pitch-duration between tones on the absolute identification of pitch. *American Journal of Psychology, 67,* 1–14.

Hayes, W. L. (1994). *Statistics* (5th ed.). Ft. Worth, TX: Harcourt Brace.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1362–1377.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273,* 1399–1402.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category level knowledge. *Science, 262,* 1747–1749.

Knowlton, B. J., Squire, L. R., Paulsen, J. S., Swerdlow, N. R., Swenson, M., & Butters, N. (1996). Dissociations within nondeclarative memory in Huntington's disease. *Neuropsychology, 10,* 538–548.

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

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science, 10,* 477–493.

Lockhead, G. R. (1970). Identification and the form of multidimensional discrimination space. *Journal of Experimental Psychology, 85,* 1–10.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

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

Maddox, W. T., & Filoteo, J. V. (2001). Striatal contributions to category learning: Quantitative modeling of simple linear and complex nonlinear rule learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society, 7,* 710–727.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8,* 1–17.

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, 85,* 207–238.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7,* 308–313.

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,* 87–108.

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

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

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5,* 345–369.

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. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science, 9,* 247–255.

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2002). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1,* 360–370.

Pitt, M. A., Myung, I.-J., & Zhang, S. (2003). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrielli, J. D. E. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology, 13,* 564–574.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Accoustical Society of America, 24,* 745–749.

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

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Comparing connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300.

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory, 5,* 420–428.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science, 12,* 318–322.

Rouder, J. N., & Gomez, P. (2001). Modeling serial position curves with temporal distinctiveness. *Memory, 9,* 301–311.

Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (in press). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review.*

Shepard, R. N. (1957). Stimulus and response generation: A stochastic model relating generalization to distance in a psychological space. *Psychometrika, 22,* 325–345.

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

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101,* 357–361.

Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 119–143.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research.* New York: Wiley.

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

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5,* 102–122.

Zaki, S. R., & Nosofsky, R. M. (2002). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective & Behavioral Neuroscience, 1,* 344–359.

Received July 9, 2001
Revision received June 27, 2003
Accepted August 5, 2003 ■