Journal of Experimental Psychology:
Learning, Memory, and Cognition
1999, Vol. 25, No. 5, 1083-1119

0278-7393/99/$3.00

A Model of Probabilistic Category Learning

John K. Kruschke and Mark K. Johansen
Indiana University, Bloomington

A new connectionist model (named RASHNL) accounts for many "irrational" phenomena found in nonmetric multiple-cue probability learning, wherein people learn to utilize a number of discrete-valued cues that are partially valid indicators of categorical outcomes. Phenomena accounted for include cue competition, effects of cue salience, utilization of configural information, decreased learning when information is introduced after a delay, and effects of base rates. Experiments 1 and 2 replicate previous experiments on cue competition and cue salience, and fits of the model provide parameter values for making qualitatively correct predictions for many other situations. The model also makes 2 new predictions, confirmed in Experiments 3 and 4. The model formalizes 3 explanatory principles: rapidly shifting attention with learned shifts, decreasing learning rates, and graded similarity in exemplar representation.

In everyday life, people encounter many situations in which they must learn to predict outcomes that are only imperfectly correlated with predictive cues. For example, check-out clerks at grocery stores learn to predict whether a customer is old enough to purchase alcoholic beverages on the basis of imperfectly predictive cues such as the customer's style of clothes, skin complexion, mannerisms, and spoken vocabulary. Seafarers learn to forecast foul weather on the basis of imperfectly predictive cues such as the size of waves, the shapes of the clouds, and the movements of animals. Investors learn to buy or sell stocks on the basis of imperfectly predictive cues such as grocery sales, the weather, and the behavior of other investors.

When learning to predict an outcome on the basis of cues, it would seem rational or optimal to learn about all the correlations as they actually exist in the world. Instead, it is the case that a more valid or more salient cue detracts from the learning of a less valid or less salient cue. Moreover, the inclusion of additional irrelevant cues detracts from the learning of valid cues. This *cue competition* in associative learning is evident across a variety of situations and across a number of animal species. It occurs, for example, when human learners make ratings of the strength of causal relationships between occurrences in a video game (Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993). Cue competition also occurs when rats or pigeons learn about associations between cues such as tones or lights and outcomes such as food, shock, or air puff (Wagner, 1969; Wagner, Logan, Haberlandt, & Price, 1968; Wasserman, 1974). Cue competition effects have even been found in learning by honeybees (Shapiro & Bitterman, 1998). The pervasiveness of cue competition suggests that it is a fundamental characteristic of cognition.

Such "irrational" learning is not limited to underutilization of competing relevant cues. In some situations people will positively utilize cues that are, in fact, irrelevant and undiagnostic. For example, when two outcomes have different frequencies of occurrence, that is, different base rates, then there are situations wherein people will tend to associate a nondiagnostic cue with the more common outcome (Kruschke, 1996a, Experiment 4). At the same time, people will apparently neglect the base rates of the outcomes, and a cue value that is objectively undiagnostic will be associated with the rare outcome (Gluck & Bower, 1988b). With such well-documented cases of irrational learning in laboratory experiments that are presumably not altogether unrepresentative of the natural environment, it is puzzling that humans (or any other of the many species that show these effects) have survived, indeed thrived, all these millenia.

In this article we consider cue utilization in a particular experimental framework, known as *nonmetric multiple-cue probability learning* (NMCPL). In this framework, multiple cues and outcomes have discrete values, hence the adjective *nonmetric*. A wide variety of interesting effects are manifested in this simple paradigm, so it is tractable yet broadly representative of more complicated scenarios. In this article we introduce a new model that addresses a wide spectrum of results emanating from this paradigm. No previously pro-

John K. Kruschke and Mark K. Johansen, Department of Psychology, Indiana University, Bloomington.

This research was supported in part by National Institute of Mental Health FIRST Award 1-R29-MH51572-01 and by an Indiana University Cognitive Science Program Fellowship.

For helpful comments on a previous version of this article, we thank Nathaniel Blair, Jerome Busemeyer, Stephen Edgell, Scott Ottaway, Robert Roe, Teresa Treat, and Pete Wood. For assistance administering the experiments, we thank Carrie Christian, James Cortright, Erin DeMien, Brandi DeMont, Christy Doherty, Jacob Hall, Scott Harris, Amy Hennies, Kara Jewell, In Pyo Lee, Keith Lyle, Angie McKillip, Andrew Nash, Amy O'Brien, Chrissy Petti, Heather Shelton, Danielle Thomas, Elyse Weiss, and Jesse Young.

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychology, 1101 East 10th Street, Indiana University, Bloomington, Indiana 47405-7007. Electronic mail may be sent to kruschke@indiana.edu. John K. Kruschke's World Wide Web page is at http://www.indiana.edu/~kruschke/.

posed model has been able to address all these effects. One psychological mechanism highlighted by this new model is limited-capacity attention that rapidly shifts to reduce error. On any given learning trial, the first thing that the model (and, we argue, a person) does is rapidly shift attention away from cues that cause error, toward cues that reduce error. The shifted distribution of attention is itself learned, so that on subsequent trials the appropriate cues can be better attended to. The shift of attention is both rash and rational: rash because it is rapid, and rational because it quickly reduces error and mitigates interference between previous and novel learning. Thus, the rash shift helps to achieve the rational goal of learning quickly. The model is therefore named *RASHNL,* which stands for Rapid Attention SHifts 'N' Learning.

Our goals in this article are, first, to review some of the major findings in NMCPL and to demonstrate that the new model, RASHNL, can account for these seemingly irrational behaviors, and second, to show that RASHNL makes two new predictions, confirmed by new experiments. The success of the model, which is a formal expression of informal explanatory principles, adds evidence to our claim that the "irrational" behaviors are a by-product of the need for rapid learning.

The article is organized as follows. We first survey a variety of phenomena observed in NMCPL and briefly review the failure of all previous models to account for these results. We then fit RASHNL to the data from two new experiments (Experiments 1 and 2) that largely replicate previous results, and we use the best fitting parameter values to predict a number of other effects reported in the literature. RASHNL also makes two novel predictions, the first prediction regarding an interaction of the effect of salience on a relevant cue and the effect of adding an irrelevant cue, and the second prediction regarding an effect of the salience of an irrelevant cue. We empirically verify these predictions in Experiments 3 and 4 and show that RASHNL fits the data well. The main section concludes by showing that RASHNL successfully accounts for apparent base-rate neglect and utilization of irrelevant cues. In the final discussion we consider how extensions to the model might address nonstationary environments, and we discuss cue competition effects observed when cues or outcomes are metric, as opposed to nonmetric. We conclude by suggesting that many different species have been exposed to similar environmental demands for rapid learning and that many different species have therefore evolved functionally similar adaptations: rapid attention shifts that produce "irrational" behavior.

Review of Previous Findings and Models

In reviewing previous findings, it will be helpful to refer to a concrete example of stimuli and procedure. Stimuli in some of our experiments consisted of a rectangle that could be tall or short, which contained a small, vertical line segment that could be positioned at the left or right, as shown in Figure 1. The outcome was one of two category labels. Thus, there were just four possible stimuli, and each

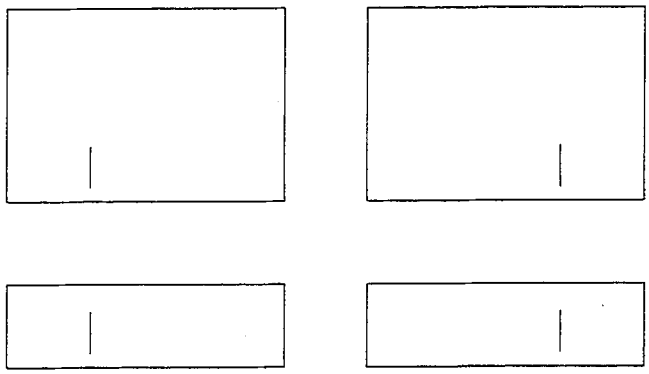

Figure 1. The four stimuli used in Experiment 1. The stimuli comprised two binary-valued cues: rectangle height (tall or short) and line segment position (left or right).

stimulus was probabilistically assigned to two categories. On any learning trial, the participant saw a stimulus, guessed the correct category by pressing the corresponding key on the computer keyboard, and was then told the correct response. The learner had to glean which category tended to go with which stimulus. Thus, NMCPL refers to a type of problem situation and is not a label for a type of cognitive process (cf. Estes, 1976).

The words *cue* and *dimension* are used synonymously in this article, although in the literature the word *cue* connotes an aspect of the physical stimulus and the word *dimension* connotes an abstract set of mutually exclusive values. The distinction is not critical for our purposes, and we hope our meaning is clear from context.

Definition of Validity and Utilization

To describe the effects found in NMCPL, we must define the terms *validity* and *utilization* precisely. Consider the example illustrated in Figure 2. In this scenario, there are two binary-valued cues, such as the cues shown previously in Figure 1. Each of the four combinations of cues occurs 10 times, for a total of 40 stimulus presentations. There are two possible categorical outcomes, J and K. The table in the top left of Figure 2 shows the frequency with which each stimulus is presented and the frequency that each stimulus is labeled as Category K by the corrective feedback. For example, the top left cell of the table, which corresponds to a stimulus with Value 1 of Cue A and Value 1 of Cue B, contains the expression "1/10," which indicates that there were 10 occurrences of this stimulus, 1 of which resulted in Outcome K (and hence 9 of which resulted in Outcome J).

The cell frequencies contain complete information about the outcomes for each stimulus, but they do not transparently indicate the correlations of the outcome with each dimension, nor the overall base rate with which each outcome occurs. We are interested in these base rates and dimensional correlations so that we can compare them with the extent to which people actually use the dimensions. The base rates and correlations can be gleaned from the marginal frequencies shown with the table, computed by summing across

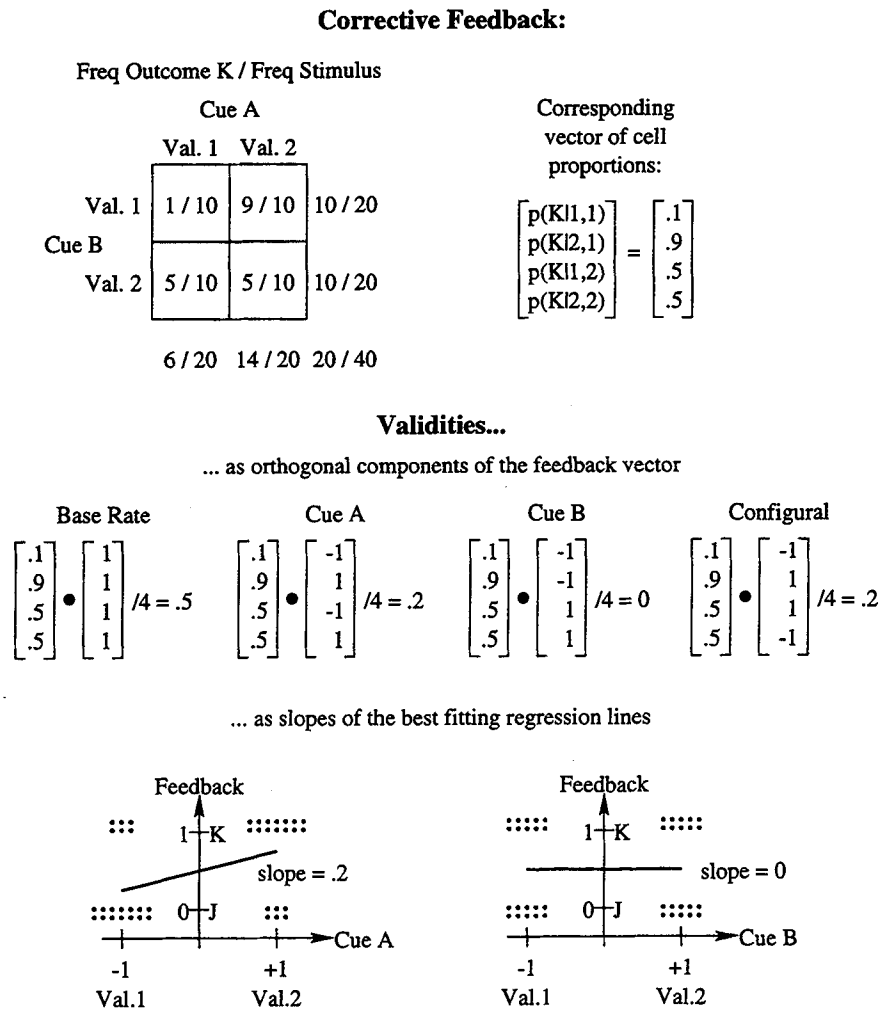

Figure 2. An example for explaining the definitions of validity (Val.) and utilization. See main text for discussion. Freq = frequency.

rows or columns. The base-rate validity expresses the overall mean probability that Outcome K occurs. This probability has a value in the range of 0–1. The cue validities express *deviations* from the overall mean probability. The example in Figure 2 shows a case in which the base rate, or overall mean probability, of Outcome K is 50% (20 out of 40 occurrences). The probability of Outcome K, conditional upon Cue A having Value 2, is 70% (14 out of 20 occurrences). This conditional probability deviates 20 percentage points from the base rate. Hence Cue A has a validity of .20.[1]

In general, the base rate and dimensional validities are computed as follows. The cell frequencies are converted to proportions and put into a column vector, shown at the top right of Figure 2. These four category proportions fully specify the structure of the feedback, assuming equal occurrence of each stimulus combination. These raw proportions do not, however, transparently indicate the correlation of the correct category with the different stimulus dimensions. The vector of raw proportions is therefore decomposed into different orthogonal components that directly reflect the dimensional influences. This decomposition is the same as statistical analysis of variance (ANOVA), which expresses factorial data in terms of a grand mean, main effects, and interaction (Edgell, 1978; Hoffman, Slovic, & Rorer, 1968).

The middle row of Figure 2 shows the raw proportion vector projected onto orthogonal vectors that encode, respectively, the grand mean or base rate of Category K, the main effect of Cue A, the main effect of Cue B, and the interaction or configural information of Cues A and B. The projection operation is accomplished via the normalized dot product, shown in the figure. For example, the second dot product

[1] Some earlier articles (e.g., Björkman, 1967; Castellan, 1973; Castellan & Edgell, 1973) defined *validity* as the product moment correlation between the cue and the outcome (i.e., the ϕ coefficient). When the various cue combinations occur with equal frequency, the validity defined in the main text is half of this correlation.

indicates that the validity of Cue A is .2. This means that there is a tendency (relative to the base rate) for the outcome to be K when Cue A has Value 2 and for the outcome to be J when Cue A has Value 1. The four validities (of base rate, Cue A, Cue B, and their configuration) are simply a re-expression of the raw proportions, such that each validity carries independent information, but which is much more transparently indicative of dimensional and configural information than the raw cell proportions.

The validities can also be conceptualized as slopes or intercepts of best fitting regression lines drawn through a scatterplot of the raw data. The lower panel of Figure 2 shows the validities of Cue A and Cue B in this manner. The values of the cues are coded as 1 and −1 on the abscissa, and the corrective feedback is coded as K = 1 and J = 0 on the ordinate. The 40 dots in each of the two graphs correspond to the 40 trials tabulated in the top left panel of the figure. For example, the lower left graph of Figure 2 shows six dots at coordinates −1, 1, indicating the six occurrences of Value 1 of Cue A that resulted in Outcome K. The best fitting regression line has a slope of .2, the same as the validity computed previously by the dot product. This graph also illustrates that the maximum possible slope (cue validity) is .5, because the range of the outcome is 1 and the range of the cue value is 2. The base rate can also be depicted graphically, as a single ordinate with no abscissa, and with the best fitting regression "line" being merely a point at the overall mean. Finally, the configural validity can be depicted analogously, but this requires some spatial gymnastics that probably would obscure more than they would reveal, so these contortions will not be demonstrated here.

The *configuration* of cues is informative when the probability of Outcome K can be better predicted by a nonlinear combination of cues than by any linear combination alone. In terms of ANOVA, the configuration is informative when an interaction is present. Expressed in terms of logical combinations of cue values, the configuration is informative when the outcome is correlated with the exclusive-or (XOR) of the cues, that is, when the mean outcome in the main-diagonal cells differs from the mean outcome in the complementary diagonal cells.

Figure 2 provides an example in which the configuration has nonzero validity. The best fitting linear combination of the marginal probabilities predicts that stimuli with Value 1 on Cue A should have a 30% (i.e., 6/20) probability of resulting in Outcome K. Contrary to this best linear prediction, the two cells for which Cue A has Value 1 do not have 30% K outcomes. Instead, for Value 1 of Cue B (and Value 1 of Cue A), the probability of Outcome K is 10% (1/10), which is 20 percentage points less than the linear prediction. For Value 2 of Cue B (and Value 1 of Cue A), the probability of Outcome K is 50% (5/10), which is 20 percentage points more than the linear prediction. Expressed in terms of an XOR of the cues, the frequency of Outcome K is 20 percentage points more than the linear prediction if Cue A has Value 2, *or* if Cue B has Value 2, but *not* if *both* cues have those values, in which case the probability is 20 percentage points less than the linear prediction. This 20-percentage-point deviation of the cell probabilities from the best linear prediction implies a configural validity of .20.

Utilization is computed the same way as validity, with the only difference being that the array now tabulates *response selections* rather than corrective feedback. For example, if a particular participant gave the same frequency of K responses as tabulated in Figure 2, then the participant's utilization would be the same as the validities, regardless of whether the participant gave the A responses on the correct trials or not.

Validities and utilizations (other than base rates) can take on positive or negative values, depending on how the "polarities" of cues and outcomes are arbitrarily coded. For example, if for Cue A, the assignment of Value 1 and Value 2 to codes −1 and 1 were reversed, then the validity and utilization would change sign.

It is important to realize that a respondent's utilizations need not necessarily have any particular relationship with the actual validities. For example, a participant might be optimal and always respond K for Value 2 of Cue A and never respond K for Value 1 of Cue A, and thereby this participant would have a Cue A utilization of .5, despite the fact that the cue's validity is only .2. A participant who responds with utilization of .5 to a cue of validity less than .5 is said to be "maximizing." Indeed, a participant might decide to utilize a cue that in fact has zero validity. In an extreme case, a perfectly contrary participant might exhibit a utilization of −.5 for a cue of .5 validity.

Summary of Effects Observed in NMCPL

The NMCPL paradigm has generated a large number of interesting effects over decades of research. Here we summarize several of the basic phenomena. All of these phenomena can be interpreted as irrational or nonoptimal.

Increased validity results in increased utilization. Learners tend to utilize a high-validity cue more than a low-validity cue, all else being equal. This has been demonstrated for *component* information (e.g., Edgell et al., 1996, Experiment 4), for *configural* information (e.g., Edgell, 1978, 1980; Edgell & Castellan, 1973), and for *base rates* (e.g., Estes, 1964). This effect might be interpreted as rational behavior, insofar as the behavior reflects the true state of the world. On the other hand, it might be interpreted as irrational insofar as optimal performance would dictate maximal utilization whenever the validity is any nonzero value. Many species exhibit this effect of approximate probability matching, or at least nonmaximizing (e.g., Mackintosh, 1970).

Increased validity of a different source results in decreased utilization. This effect is the classic cue competition phenomenon. When two sources of information are present, they tend to compete, such that increasing the validity of the second source, while leaving the validity of first source *fixed,* tends to increase the utilization of the second source while *decreasing* the utilization of the first, fixed-validity source. This competition effect has been documented for cases in which the fixed-validity source is a component and the competing source is a component

(Edgell et al., 1996, Exp. 4) and when the competing source is configural (Edgell, 1978, 1980, 1993; Edgell & Castellan, 1973; Edgell & Roe, 1995), although in the latter cases the competitive trends sometimes did not reach statistical significance. From a rational or normative perspective, when cues occur independently, the one of highest validity should be utilized maximally, regardless of the validities of the other cues. As mentioned earlier, cue competition effects have been demonstrated in a variety of experimental settings and in a range of species. In learning paradigms, cue use has been shown to depend inversely on the validity of other cues, for humans, rats, and pigeons (e.g., Baker et al., 1993; Wagner, 1969; Wagner et al., 1968; Wasserman, 1974).

Increased salience results in increased utilization. When the salience of information is increased, it tends to be utilized more. The salience of a cue can be defined in various ways (e.g., as the discriminability or dissimilarity of two alternative values of the cue). The effect of salience has been well documented for *component* information (Edgell, Bright, Ng, Noonan, & Ford, 1992; Edgell et al., 1996, Experiments 4 and 6). Comparable effects have been observed in concept learning experiments (which use deterministic mappings from stimuli to categories), discrimination learning (in which two stimuli are presented simultaneously), and classical conditioning (in which the categorical outcome is the unconditioned stimulus). For a useful review, see Trabasso and Bower (1968, chap. 6).

It is not yet clear how to define the salience of *configural* information, however. For highly separable component dimensions, the salience of their configuration should, presumably, depend only on the saliences of the components. For highly integral component dimensions, the salience of a configuration should be quite distinct from the salience of the components.

Increased utilization of more salient information is reasonable from an intuitive perspective, in that people will, by definition, better notice and learn about more attractive stimuli. From a rational perspective, however, any information of nonzero validity should be maximally utilized, whether it is salient or not.

Increased salience of a different source results in decreased utilization. When two sources of information are present, increasing the salience of one source tends to decrease the utilization of the other source. In conditioning paradigms, a similar effect is referred to as *reciprocal overshadowing* and has been observed in both humans and rats (e.g., Mackintosh, 1976; March, Chamizo, & Mackintosh, 1992; Ruebeling, 1993). A related effect has also been observed, anecdotally, in hunting dogs, who can be distracted from pursuit of their prey by a red herring dragged across their path. The effect is so strong that it has become a colloquialism.

This competitive effect of salience has *not,* however, been verified in NMCPL by Edgell and colleagues when the competing cue has zero validity. For example, Edgell et al. (1992, Experiments 4 and 5) found a nonsignificant trend toward this competitive trade-off and concluded that the null effect was true (Edgell et al., 1992, p. 587). Edgell et al. (1996, Experiments 5 and 6, p. 1477) also found some nonsignificant trends and concluded that no competitive effect occurred; that is, the salience of irrelevant dimensions does not affect the utilization of a valid dimension.

It is possible that the conditioning paradigm, in which a competitive effect of salience has been found, and the NMCPL paradigm, in which Edgell and colleagues have not found a significant competitive effect of salience, differ in some critical way that erases salience-based competition in NMCPL. It is also possible, however, that the nonsignificance of the trends observed by Edgell et al. was a consequence of experiments with low statistical power, because, as will be seen, the within-condition variance in these experiments is large and the effects of salience can be small.

There are reasons to expect that the competitive effect of salience should occur in NMCPL. First, there are numerous analogies between other effects observed in conditioning and effects observed in NMCPL. Second, consider the extreme case of changing the salience of a competing dimension from *zero* to some typical nonzero value. That is, the competing dimension is either absent (zero salience) or present (nonzero salience). This change has a robust effect on utilization, even when the added source has zero validity, as described in the next section. Logically, then, comparable changes in salience, from a small nonzero value to a large nonzero value, should also have an effect. If no such effect occurs, then there is a qualitative difference between changes in salience that do not affect presence and changes in salience that do affect presence. A third reason to expect a competitive effect of salience is that RASHNL strongly predicts that it should occur. Experiment 4 of this article tests this new prediction.

The addition of an irrelevant source results in decreased utilization. When an irrelevant component is added to a stimulus, then the utilization of a fixed-validity source tends to decrease. This competitive effect has been thoroughly documented in a number of studies. Castellan (1973) compared conditions in which a *component* of fixed validity was accompanied by different numbers of irrelevant cues. He found that utilization of the valid component decreased as the number of irrelevant cues increased. The magnitude of the decrease was strongest for middling validities. Edgell and Hennessey (1980) found an analogous effect for the utilization of *base rates.* As the number of irrelevant dimensions was increased, the utilization of the base rates decreased, and the decrease was strongest for middling base rates. Edgell et al. (1996, Experiments 1 and 2) reported similar effects for *configural* cues. The utilization of a relevant configuration decreased when irrelevant cues were added. In reasoning or problem-solving paradigms, it is also the case that cue utilization depends inversely on the presence of other irrelevant cues (e.g., Goldstein, 1973; Nisbett, Zukier, & Lemley, 1981). Thus, cue competition in NMCPL is representative of a pervasive phenomenon in cognition.

The deleterious influence of an added irrelevant source depends on its salience. When an irrelevant cue is added to a relevant cue of greater salience, then the utilization of the

salient cue will decrease relatively little. When the same irrelevant cue is added to a relevant cue of lesser salience, however, then the utilization of the relevant cue will decrease noticeably. This interaction is a new prediction of RASHNL, tested in Experiment 3 of this article.

Delayed introduction of information results in retarded learning. Probabilistic classification presents a very difficult environment to the learner, because no matter how much effort the learner devotes to the task, even optimal responding generates a large amount of error. The learner must resign herself or himself to this unavoidable error. It is plausible that in such situations, people quickly learn to discount their errors and reduce the rate at which they alter their knowledge; that is, learners reduce their learning rates after a limited number of trials. Busemeyer and Myung (1988) reported evidence of learners decreasing their learning rates when the participants' task was to learn the central tendency of a probabilistic distribution. Indeed, many computational learning algorithms are designed to *anneal* (i.e., decrease) their learning rates in probabilistic environments (e.g., Almeida, Langlois, Amaral, & Plakhov, 1998; Darken & Moody, 1991, 1992; Sompolinsky, Barkai, & Seung, 1995), and convergence proofs for various learning algorithms assume decreasing learning rates (e.g., White, 1989).

Edgell (1983) and Edgell and Morrissey (1987) reported results consistent with this reduction of learning rates. Training in their experiments began with certain components or configurations being informative, but with other information being introduced after a delay. For example, some experiments began with just one cue having a nonzero validity, but after a number of trials, the configural validity changed from zero to a positive value. The experiments showed that configural information, introduced only 20 trials after the beginning of the experiment, was utilized far less than when the same information was available at the beginning of learning. Dimensional (as opposed to configural) information was also utilized less if its introduction was delayed, with the decrement in utilization being larger for larger delays.

Utilization of component information is greater than utilization of configural information. Several experiments have shown that configural information can be learned, and, for components and configurations of equal validity, the component information will tend to be utilized more than the configural information, as long as the components are of comparable salience (Edgell, 1980, 1993; Edgell et al., 1996; Edgell & Roe, 1995). This statement also presumes that the components are highly separable dimensions. Utilization of component and configural information has not been compared for highly integral dimensions.

An irrelevant component can be positively utilized. When a cue is strongly but equally correlated with both of two outcomes, and one outcome occurs with a much higher base rate than the other, then people will tend to learn that the cue is a predictor of the more common outcome. Such learned utilization of an irrelevant cue has been documented in a disease diagnosis situation (Kruschke, 1996a, Experiment 4), and the phenomenon poses a difficult challenge for models of associative learning.

Apparent base-rate neglect: An undiagnostic cue value can elicit unequal preferences. Another celebrated phenomenon observed in NMCPL is apparent base-rate neglect (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988b; Kruschke, 1996a; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990, 1991a, 1991b). This effect is observed in experimental designs that have been carefully constructed such that a particular cue value, V, occurs more often for a rare outcome, R, than for a common outcome, C, but the probability of the rare outcome, given the cue value alone, is the same as the probability of the common outcome. Formally, $p(V|R) > p(V|C)$, but $p(R|V) = p(C|V)$. The statistics of this experimental design dictate that a rational learner, when tested with cue value V alone, should respond equivocally with outcomes R and C. In fact, there is a strong tendency for people to prefer the rare outcome R over the common outcome C. This effect has also been an important challenge for recent models of learning (Kruschke, 1992, 1996a; Lewandowsky, 1995).

Previous Theories and Models

The various phenomena, outlined above, constitute a formidable challenge for models of learning. In this section we review a number of previously proposed models and describe how all of them fail to account for one or more of the effects.

Hypothesis-generation model. Castellan and Edgell (1973) proposed to account for results from NMCPL with two models that generate hypotheses about which sources of information lead to correct answers and hypotheses about which responses are appropriate for the attended-to sources. In one version of the model, only the base rates and component dimensions can be attended to. In a second version, configurations of cues can also be attended to. The first version, of course, cannot possibly learn to utilize configural information, whereas people can. The second version, which can learn configural information, unfortunately makes several incorrect predictions. For example, Edgell et al. (1996, p. 1466) reported that it predicts no decrease in utilization of a relevant dimension when irrelevant dimensions are added. Edgell and Roe (1995) reported other shortcomings of the model, regarding ordinally incorrect predictions about relative utilization of configural and dimension information.

Rule-plus-exception model (RULEX). More recently, Nosofsky, Palmeri, and McKinley (1994) proposed another model, named *RULEX,* in which hypotheses (i.e., rules) are generated. In this model, learning consists of first searching for perfect single-dimensional rules, then, if no such rules exist, searching for imperfect single-dimension rules, then, if no such rules exist, searching for conjunctive rules. If success is found at any stage, then exceptions to rules are also added to limited-capacity memory. The initial form of RULEX was designed for discrete-valued stimuli and deterministic mappings to categories. A later form of RULEX, described by Nosofsky and Palmeri (1998), applies to continuous-valued stimuli but was not thoroughly developed for probabilistic mappings. Thus, it is not clear whether

RULEX in its current forms can be sensibly applied to NMCPL. Assuming that some modification could be made so that RULEX could generate probabilistic responses, along the lines suggested by Nosofsky and Palmeri (1998), it remains doubtful that the model would show decreased utilization of a valid dimension when irrelevant dimensions are added. Resolution of this matter awaits future development of RULEX for application to NMCPL.

Contingency (ΔP), probabilistic contrast, and Power PC models. Models of causal induction address how people infer the extent of causal relationship between candidate cause c and effect e. Basic contingency theories assert that causal strength corresponds to the *contingency* between the candidate cause and the effect, defined as $\Delta P_c = P(e|c) - P(e|\bar{c})$, where $P(e|c)$ is the observable probability that the effect occurs given that the cause occurs, and $P(e|\bar{c})$ is the observable probability that the effect occurs given that the cause does not occur. (Notice that ΔP equals twice the validity as defined for NMCPL.) The simplest form of this scheme is inadequate for addressing situations with multiple, covarying candidate causes. Cheng and Novick (1990, 1992) introduced an extension of the contingency approach by arguing that causal strength corresponds to the contingency computed when conditionalized on certain focal sets of events. In this *probabilistic contrast model (PCM)*, the causal strength corresponds to $\Delta P_{c,F} = P(e|c, F) - P(e|\bar{c}, F)$, where F denotes the focal set. The theory says that reasoners prefer focal sets in which plausible alternative causes are controlled, that is, held constant (either present or absent). When the PCM is applied to the design shown in Figure 3, the contingency of the .2-validity cue is the same regardless of the focal set used to compute it. If utilization is monotonically related to inferred causal strength, then it is not clear how the PCM can account for a reduction in utilization either when the validity of uncorrelated dimensions is increased or when irrelevant dimensions are added.

Cheng (1997) significantly modified this approach by arguing that the inferred causal power, p_c, of a candidate cause, c, is related to observable contingencies indirectly via combined influences with other potential causal powers. When candidate causal factor c occurs independently of all other candidate causal factors, the *Power PC* theory predicts that inferred causal power will be ordinally related to $p_{c,F} = [P(e|c, F) - P(e|\bar{c}, F)]/[1 - P(e|\bar{c}, F)]$ (cf. Cheng, 1997, Equation 8, p. 374). This theory significantly enhances the scope of the contingency approach, but still the theory appears unable to account for reductions in utilization when independent, irrelevant cues are added, because this addition does not affect any of the terms that influence causal power.

Context model. The context model, proposed by Medin and Schaffer (1978), stores memory traces of stimulus instances with their categorical labels. When a stimulus is presented to the model, activation of memory exemplars is based on their similarity to the stimulus. The probability of a particular categorical response is based on the summed activation of all exemplars of that category, relative to the summed activation of all known exemplars. Edgell et al. (1996, p. 1466) reported that the context model predicts *no* decrease in utilization of a relevant dimension when irrelevant dimensions are added, because the relative proportions of activated exemplars remains the same.

Generalized context model (GCM). Medin and Schaffer's (1978) context model was generalized by Nosofsky (1986). The GCM added an assumption that when computing the similarity of a stimulus to a memory exemplar, the contribution of each dimension was weighted by the attention allocated to that dimension, and the attention strengths were constrained to sum to one. This capacity constraint on attention suggests that the GCM might be able to account for competition between cues in some situations. The GCM also assumed, in lieu of a mechanism for adjusting attention through learning, that attention is distributed across cues *optimally.* In particular, this implies that irrelevant cues receive zero attention. Hence, the GCM incorrectly predicts that the addition of irrelevant cues will not decrease the utilization of a relevant cue.

Rational model. Anderson (1990, chap. 3, 1991) proposed a category learning model in which instances can be clustered together, and a new instance is classified according to (a) the probability with which the instance belongs to the existing clusters and (b) the probability of the classifications for each cluster. When each cluster contains just one instance, the rational model reduces to the context model (Nosofsky, 1991). The rational model is motivated from normative Bayesian statistics, with implementational constraints (hence its appellation). Because the rational model is able to form clusters of instances, it might be able to address phenomena that the context model cannot. Anderson (1990, pp. 120–125) applied the rational model to apparent base-rate neglect and reported that the model could not produce this effect. This discrepancy was dismissed, at the time, by saying that participants in the Gluck and Bower (1988b) experiment, who were asked to rate the probability of a disease given a symptom, were mistakenly replying with their rating of the symptom given the disease. The effect has been replicated several times, however, using category choice instead of probability rating, and so this "irrational" phenomenon cannot be explained as mere failure of the participants to follow instructions.

	.0, .2		**.2, .2**		**.3, .2**	
	Cue A		Cue A		Cue A	
	Val. 1	Val. 2	Val. 1	Val. 2	Val. 1	Val. 2
Val. 1	3 / 10	3 / 10	1 / 10	5 / 10	0 / 10	6 / 10
Cue B						
Val. 2	7 / 10	7 / 10	5 / 10	9 / 10	4 / 10	10 / 10

Figure 3. The three conditions of Experiment 1. Each table indicates the distribution of "category K" feedback for a block of 40 trials. In all three conditions, the validity (Val.) of Cue B was .2. Cue A had validities of .0, .2, and .3 in the three conditions. From these tables it can also be determined that the base rates were fixed at .5 across all three conditions, and the configural validity was fixed at 0 across all three conditions.

Component-cue, configural-cue, and ALCOVE connectionist models. The component-cue model of Gluck and Bower (1988b) is a linear connectionist network that builds associative weights between input cues and output categories proportionally to the error between actual and predicted classifications. The model made a significant impact, because it was a simple generalization of the Rescorla and Wagner (1972) model of animal conditioning and because it accounted for apparent base-rate neglect. Unfortunately, it has since been shown to fail to account for utilization of an irrelevant cue (Kruschke, 1996a, Experiment 4), in an experimental design very similar to the original designs that demonstrated apparent base-rate neglect. Edgell, Roe, and Zurada (1993) showed that the model also fails to account for the detrimental effect of additional irrelevant cues, as does the configural-cue model of Gluck and Bower (1988a), in which not only component-cue values but also their configurations are represented at the input to the network.

The ALCOVE [Attention Learning COVEring] model of category learning (Kruschke, 1992) adds a connectionist learning mechanism to the GCM. ALCOVE was originally thought to exhibit apparent base-rate neglect, but this result was later shown to be severely restricted to particular training sequences (Lewandowsky, 1995). Busemeyer, Myung, and McDaniel (1993b) showed that a large class of error-driven learning models—including ALCOVE, the component-cue, and configural-cue models—will converge, at asymptote, to the same utilization of a valid dimension despite differences in the number of irrelevant dimensions. Thus, these connectionist models also fail to account for at least some of the "irrational" phenomena compiled above.

Conclusion regarding previous models. Many of these previous models lack a quality that is assumed by essentially all theories of cue competition effects, namely, limited capacity attention. Because irrelevant dimensions are, by chance, correct on a subset of trials, attention will be diverted away from valid dimensions to the irrelevant dimensions, thereby reducing the utilization of the valid dimensions. This notion of competition for limited attentional capacity also permeates explanations of cue-competition effects in other paradigms (e.g., Sutherland & Mackintosh, 1971; Trabasso & Bower, 1968). Despite the pervasiveness of this notion of limited attention, it has not previously been formalized in a successful model of NMCPL. Previous models of hypothesis testing can be interpreted as using rapid shifts of attention, insofar as hypotheses that put conditions on different dimensions are attending to different dimensions. When taken to an extreme, the rapid attention shifts in RASHNL might be interpreted as changes in hypotheses. Limited-capacity attention, with rapid shifts, is a central characteristic of the new model we present in this article.

In addition to incorporating competition for attention, Edgell et al. (1996) and Edgell et al. (1992) argued that any successful model will have to incorporate a memory error mechanism whereby less salient cues are more prone to errors in memory. It is this memory error mechanism that is purported to underlie effects of salience on utilization. The theory suggests that lower salience cues are more prone to confusion in short-term memory (STM). Because the values of the cues in STM have been made more random by errors of confusion, the effective validity of cues in STM is less than the physical validity. Because learning operates on the contents of STM, the less salient cues are utilized less than the more salient cues. The new model we present in this article can be interpreted as implementing this notion of blurred cue values in STM. What mediates the blurring is similarity-based exemplar representation. The results of our Experiment 4 suggest a modification to Edgell et al.'s theory, however.

None of the previous models applied to NMCPL has explicitly addressed learned nonlearning, that is, the rapid deceleration of learning over the initial training trials. This change of learning rate is the third critical principle of the new model introduced in this article. Although the model we present is a connectionist model, it does not suffer the same failures as the other connectionist models reviewed above.

Experiment 1: Effects of Competing-Cue Validity

The primary purpose of this experiment was to generate data that show a decrease in utilization of a fixed-validity cue when another dimension's validity increases. A secondary purpose was to show the effect of cue salience on utilization. These data can then be used for parameter estimation in RASHNL, which in turn can be used to make predictions for other situations.

Our stimuli, as shown in Figure 1, were rectangles that had two possible heights and that contained a small vertical line segment that had two possible lateral positions. Previous research in our lab has demonstrated that the lateral position of the line segment is more salient than the height of the rectangle (e.g., Erickson & Kruschke, 1998, Appendix C).

Our design included three conditions in which the two dimensions had various validities. In every condition, one of the dimensions had a validity of .2. The other dimension had a validity of 0, .2, or .3 across the three conditions, which are therefore referred to as the ".0, .2"; ".2, .2"; and ".3, .2" conditions, respectively. Figure 3 shows how this design was realized in terms of frequencies of feedback for particular stimuli. On the basis of previous results in the literature, we anticipated finding a reduction in utilization of the .2-validity dimension as the validity of the other dimension increased. We also expected to find greater utilization of the more salient dimension (i.e., the line segment) than of the less salient dimension (i.e., the rectangle height) when they had equal validities.

Edgell et al. (1996, Experiment 4) showed that utilization declined as the validity of a second dimension increased. We extend their experiment in two modest ways. First, they tested conditions with dimensional validities of .0, .2; .1, .2; and .2, .2, so that the validity of the fixed-validity cue was never exceeded by the other cue, whereas one of our conditions has validities of .3, .2. Second, they had only a single assignment of physical cues to abstract dimensions, wherein the somewhat less salient cue was assigned to the fixed-validity dimension. We measured utilization when the

fixed-validity cue was the less salient cue but also when it was the more salient cue. This manipulation allowed us to measure an effect of salience on the magnitude of cue competition.

Method

Participants. A total of 271 students volunteered for partial course credit in an introductory psychology class at Indiana University.

Stimuli and apparatus. Participants were trained individually in dimly lit, sound-dampened cubicles. They sat before an IBM-compatible PC at a comfortable viewing distance. They made responses by pressing the "F" or "J" keys on the standard computer keyboard.

Stimuli, illustrated in Figure 1, were presented on the computer monitor. The rectangle width was 98 mm, and its two possible heights were 32 and 68 mm. The interior line segment had a height of 15 mm, had alternative positions separated by 35 mm, and appeared 4 mm above the lower line of the rectangle. The lines were yellow against a black background.

Design and procedure. The three validity conditions were described above, as shown in Figure 3. Each participant was trained for 10 blocks of 40 trials. Trials were randomly permuted within each block. This number of trials replicates the quantity used in previous experiments by Edgell and colleagues, who were interested, in part, in approximately asymptotic utilization.

The assignment of physical stimulus dimensions to abstract cues was counterbalanced across participants. Thus, rectangle height could be assigned to Cue A or to Cue B, with line position being assigned to the other cue. The polarity of each dimensional assignment was also counterbalanced, so that, for example, if rectangle height was assigned to Cue A, then Value 1 could be the tall rectangle (for half the participants) or the short rectangle (for the other half of the participants). The assignment of "F" and "J" response keys to "K" and "J" categories was also counterbalanced.

Participants were rotated through the three validity conditions and 16 stimulus realizations in the presumably random order that they volunteered for the experiment, spread over many months. The various conditions of this experiment were conceived of in several different phases, and some of the conditions were rotated with other conditions not relevant here. Therefore, the different conditions had different numbers of participants. Condition .0, .2 had a total of 32 participants with rectangle assigned to the .2-validity cue and 63 participants with rectangle assigned to the .0-validity cue. The latter case ended up being just 1 participant short of a complete counterbalance, because we ran out of participants at the end of the semester, but, as will be seen below, the variance across participants is so large that this incomplete counterbalance should have essentially no influence on the conclusions from the results. Condition .2, .2 had a total of 112 participants. The .3, .2 condition had 32 participants with rectangle assigned to the .3-validity cue and 32 participants with line segment assigned to the .3-validity cue.

Instructions were presented on the computer screen. Previous work with probabilistic category learning indicated that participants find probabilistic relationships to be very frustrating and difficult to learn, because of the impossibility of error-free performance. Therefore, the instructions emphasized that the participant's goal was to learn an imperfect tendency of certain outcomes for certain stimuli and that if the participant tried hard, she or he could get up to 70 or 80% correct. The instructions mentioned nothing about either cue being more or less informative than the other. By contrast, Busemeyer, Myung, and McDaniel (1993a, Footnote 5, p. 194) found it necessary, in their metric cue learning task, to instruct learners that one of the cues was more effective than the other, although learners were not told which cue was more effective. Without this instruction, these researchers observed cue cooperation instead of cue competition. The complete text of our instructions is provided in Appendix A.

On each trial of training, a stimulus appeared, with a prompt below it that read, "Is this an F or a J?" When the participant made a response, the stimulus stayed on the screen, but the prompt was replaced with corrective feedback that indicated whether the response was correct or wrong, and the correct label. Wrong responses were also followed by a brief tone. The participant could study the stimulus and feedback for up to 30 s and pressed the space bar to proceed to the next trial. In addition to trial-by-trial corrective feedback, the computer displayed the participant's percentage correct for the previous block at the end of each block of 40 trials.

Results and Discussion

Figure 4 shows the mean utilizations of each cue as a function of the experimental condition and block of training.

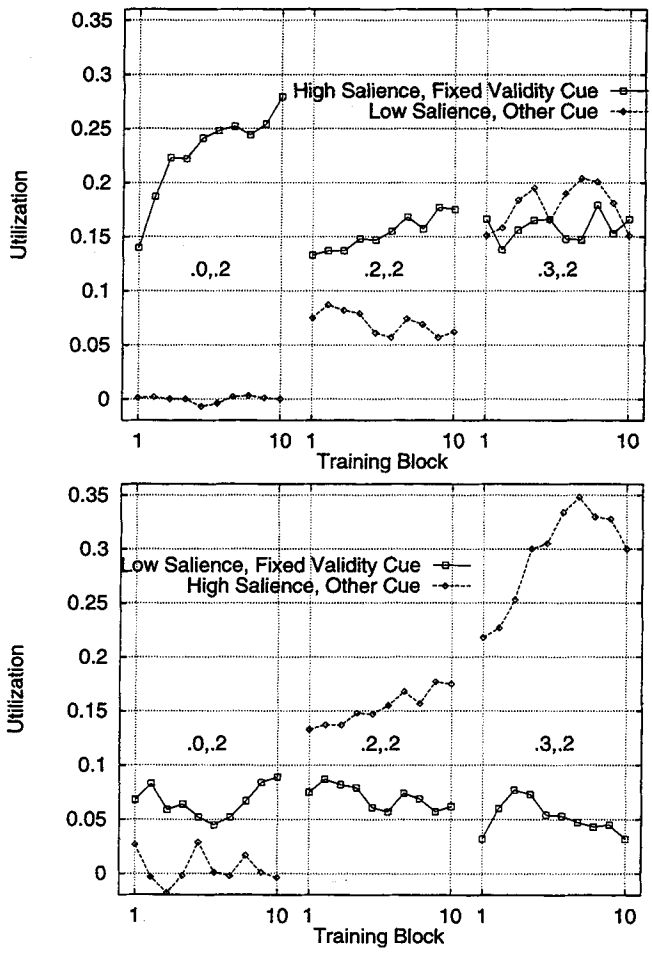

Figure 4. Results of Experiment 1. The abscissa is subdivided into three recurrences of training Blocks 1–10, corresponding to the three different validity conditions (.0, .2; .2, .2; and .3, .2, respectively).

The upper panel shows the conditions for which the fixed .2-validity cue was the high-salience line position. The lower panel shows the conditions for which the fixed .2-validity cue was the low-salience rectangle height. The data for the .2, .2 conditions are the same in the upper and lower panels, with the roles of "fixed validity cue" and "other cue" reversed.

One striking aspect of these learning curves is that utilizations are already at a fairly high level within the first block of training, and utilizations rise relatively little after the first three blocks of training. The changes from one block to the next are slight compared with the variance within blocks (described below). These data, and informal comments from participants, indicate that people learn whatever they can in the first few dozen trials and then mostly merely maintain this pattern of responding. This interpretation is consistent with the results, summarized earlier, of Edgell (1983) and Edgell and Morrissey (1987), who found that when a cue's validity was increased from zero to a moderate value after delay, the cue was not utilized as quickly as when it was valid from the beginning of training.

We are primarily interested in the effect of other-cue validity on utilization of a fixed-validity cue, and we are also interested in the effect of cue salience on the utilization of a cue. We therefore consider the utilization of the .2-validity cue, when it was either the high-salience line segment position or the low-salience rectangle height, and when it was combined with another dimension of .0, .2, or .3 validity. Data from these conditions are plotted with squares in Figure 4. Each participant's mean utilization of the dimensions was collapsed across Blocks 4–10 (as plotted and discussed in Figure 5). This 2 (salience) × 3 (validity of other cue) design has the two .2-validity cells using (different) data from the same participants, thereby making this a mixed between- and within-subjects design. For simplicity, we treated the data as completely between subjects. Hence the N for analysis purposes was 271 plus 112, which totals 383. The utilization of the fixed-validity dimension did indeed decrease as the validity of the other dimension increased, $F(2, 377) = 6.43$, $MSE = 0.111$, $p = .002$. The utilization of the more salient line segment was much greater than the utilization of the less salient rectangle height, $F(1, 377) = 75.5$, $MSE = 1.30$, $p < .001$. The greater utilization of the more salient dimension is also evident within the .2, .2

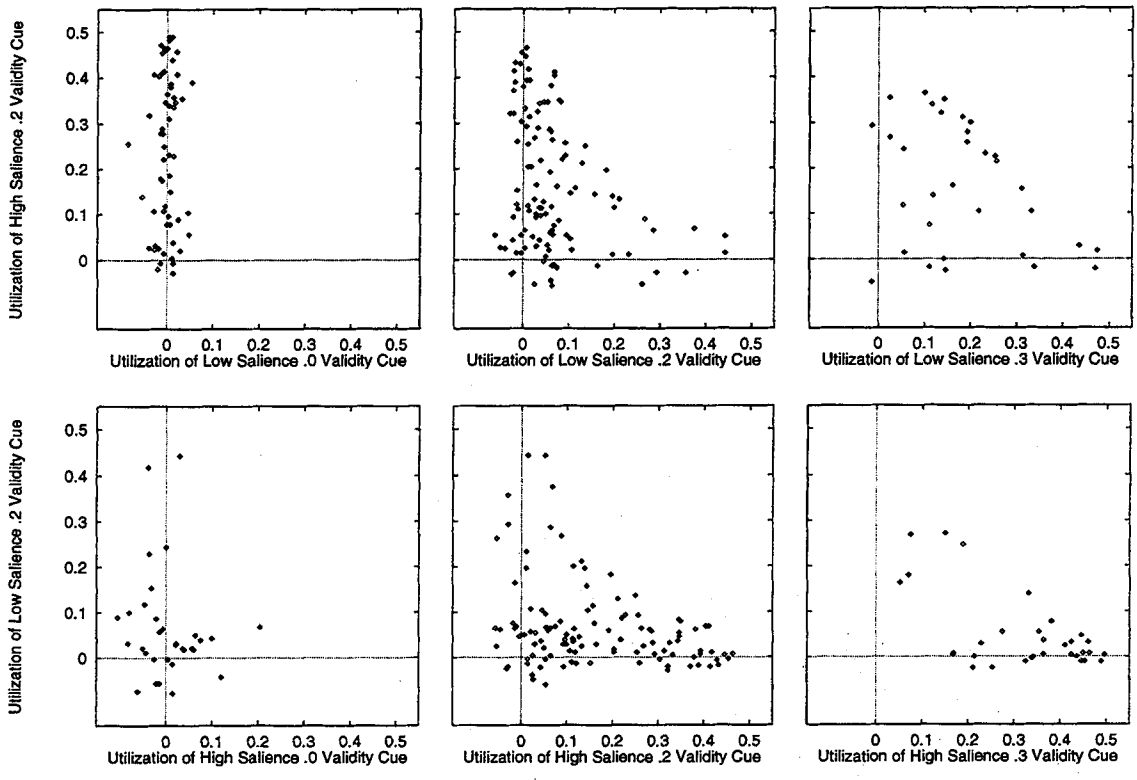

Figure 5. Utilizations of the two cues by individual participants in Experiment 1. Each point in a graph shows the mean utilizations of the two cues by a single participant, collapsed across Blocks 4–10 of training. The upper three graphs of this figure correspond to the three conditions in the upper panel of Figure 4, and the lower three graphs correspond to the lower panel of Figure 4. The two middle graphs show the same data, but with the roles of the axes exchanged. As a consequence of the design constraint that all cue combinations have an equal frequency, participants cannot have utilizations of .5 on both cues. In fact, the sum of the utilization of Cue A (UA) and the utilization of Cue B (UB) must be less than or equal to .5. Thus, points on the graph must lie within the diamond-shaped region defined by $|UA| + |UB| = .5$.

condition, $t(111) = 5.05$, $SE = 0.019$, $p < .001$. There was also an interaction of other-cue validity and salience, $F(2, 377) = 3.52$, $MSE = 0.0607$, $p = .031$, such that the effect of other-cue validity was due primarily to the high utilization of the line segment in the .0, .2 condition, which can be seen in the upper left of Figure 4.

The distributions of utilizations within conditions were noticeably nonnormal, and so, as a check on the statistical inferences described above, the data were transformed various ways in an attempt to better approximate normal distributions. Various continuous transformations yielded statistical conclusions the same as those reported above. An ANOVA was also conducted on the ranks of the utilizations (Conover & Iman, 1981). With this less powerful analysis, the two main effects were again highly significant, but the interaction failed to reach significance, $F(2, 377) = 2.22$, $MSE = 22{,}183.8$, $p = .110$.

As an additional check on the statistical conclusions, the .2, .2 group of 112 participants was split into two random halves of 56 participants, with each half contributing only to the low-salience or high-salience condition. Thus, the design was purely between subjects. An ANOVA showed that the main effects were highly significant, $F(2, 265) = 5.40$, $MSE = 0.094$, $p = .005$, and $F(1, 265) = 74.3$, $MSE = 1.29$, $p < .001$, but the interaction did not reach significance, $F(2, 265) = 1.72$, $MSE = 0.030$, $p = .182$.

Our results qualitatively reproduced those of Edgell et al. (1996, Experiment 4) for the conditions that we replicated (shown in Figure 4 as the left and middle partitions of the upper panel). The interaction of other-cue validity and salience is a new result, but its statistical reliability is marginal in the present data, due to the large within-group variance, described next.

Figure 5 shows the range of individual utilizations for the conditions in Experiment 1. Each data point represents the mean cue utilizations by a single participant, collapsed over Blocks 4–10, that is, the last 280 trials. One striking aspect of the individual utilizations is their range: Some participants barely utilized either cue and may as well have had their eyes closed, whereas other participants virtually maximized utilization of one cue or the other. Still other participants split their utilizations across both cues.

The fact that some participants showed nearly zero utilization of both dimensions might indicate that these participants were not following instructions and that their data should be excluded from our analysis. Unfortunately, no clear criteria exist for determining which data should be excluded. Edgell et al. (1996) omitted data if they exhibited long runs of the same response, which might indicate a participant's purposeful neglect of learning. For example, in their Experiment 4, data from 51 of 279 participants were discarded. There is, however, no clear criterion for how long a run needs to be for it to indicate neglect, and there is no method for establishing what other types of patterns do or do not indicate intentional negligence (Edgell et al., 1996, Footnote 1, p. 1473). Therefore, to be conservative, we did not exclude any participants' data from our analysis. Even if we did exclude participants who gave long strings of the same response, and who therefore utilized neither dimension, we would retain participants who did utilize one or both dimensions, and the large individual differences in our data would persist.

The individual differences seen in our data reflect those observed by Trabasso and Bower (1968, chap. 3). In their concept learning experiments, stimuli had *redundant relevant cues,* such that both of two cues had a validity of .5. For example, a triangle with a dot below it would always indicate one category, whereas a circle with a dot above it would always indicate the other category. (Conflicting cue combinations, such as a triangle with a dot above it, or a circle with a dot below it, never appeared during training.) The relative saliences of dot position and shape were assayed by separate experiments, which indicated that dot position was more salient than shape. Utilizations of the redundant relevant cues were assessed in a sequence of test trials during which only one of the cues was presented, without corrective feedback. Trabasso and Bower (1968, p. 78) found that 50% of their 89 participants utilized only the more salient dot-position cue, 35% of the participants utilized only the less salient shape cue, and the remaining 15% utilized both cues. Thus, the type of individual differences observed in our experiments are robust and generalize from probabilistic to deterministic learning paradigms.

Edgell et al. (1996) also mentioned the occurrence of large variances of individual utilizations in NMCPL experiments, but no previous reports in the literature have displayed the distribution of individual differences. Whereas all previous work emphasized group mean utilizations, future work will have to address these large individual differences.

Experiment 2: Effects of Salience

As reviewed above, previous research has shown that high-salience cues are utilized more than low-salience cues of the same validity. This phenomenon is known as *overshadowing* in the animal learning literature. Overshadowing was seen in the .2, .2 condition of Experiment 1, wherein dimensions of equal validity had unequal utilization, corresponding to their unequal salience. Edgell and colleagues (Edgell et al., 1992, 1996) have also shown that the utilization of a cue depends on its salience even if the other cue has zero validity. Experiment 1 replicated this result in the .0, .2 conditions, wherein the .2-validity line segment position was utilized much more than the .2-validity rectangle height. In both these cases, however, we were comparing the utilizations of different cues. In Experiment 2, we instead manipulated the salience of a single dimension while we leave the other dimensional salience constant.

Experiment 2 uses two variations of the .0, .2 condition of Experiment 1. In both variations, the rectangle height has a validity of .2, and the line segment has a validity of zero. In one variation, the difference in height between tall and short rectangles is smaller, hence less salient, than in Experiment 1. In the second variation, however, the difference in height between tall and short rectangles is larger, hence more salient, than in Experiment 1. Previous authors have found that the magnitude of the difference affects utilization, and

they have explained this effect in terms of the physical difference corresponding to a psychological salience (Edgell et al., 1992; Trabasso & Bower, 1968). We expected to find a difference in utilization corresponding to the saliences of the rectangle.

Method

Participants. Students (n = 112) volunteered for partial course credit in an introductory psychology class at Indiana University.

Stimuli and apparatus. The apparatus was the same as that used in Experiment 1. The line segment size and positions, and the rectangle width, were the same as in Experiment 1, but the rectangle heights were different. In the high-salience condition, the two possible heights were 23 and 108 mm. In the low-salience condition, the two possible heights were 53 and 71 mm. The lines were inadvertently made slightly thicker than in Experiment 1, approximately 2 mm thick (as opposed to approximately 1 mm thick in the previous experiment). We presume that this small change in thickness had negligible effects on the results.

Design and procedure. The procedure was the same as for Experiment 1. Participants were assigned to the conditions by simple rotation, in the presumably quasi-random order in which they arrived at the lab; hence, there were 56 participants per condition. The assignment of abstract values to physical values was again fully counterbalanced.

Results and Discussion

Learning curves for the two conditions are shown in Figure 6. As in Experiment 1, utilization was fairly high already in the initial blocks of training, with little subsequent change in Blocks 4–10. Collapsing across Blocks 4–10, the mean utilizations of the high- and low-salience rectangles were .110 and .051, respectively. This difference was statistically significant, $t(110) = 2.50$, $MSE = 0.016$, $p = .014$. (The difference remained highly significant when the data were transformed to achieve more normal distributions.) The distributions of individual utilizations were qualitatively comparable to the left panels of Figure 5, so graphical displays are not included here.

In Experiment 1, the mean utilization of the rectangle in the .0, .2 condition was .065, which lies between the utilizations observed in Experiment 2. This is appropriate insofar as the height variation in Experiment 1 was also intermediate between the high- and low-salience height variations of Experiment 2.

Experiment 2 has therefore replicated the finding that utilization of a dimension depends on the magnitude of variation on that dimension, when it is in the presence of another dimension of zero validity. Experiment 2 provides us with further data for quantitative model fitting.

Summary of Results From Experiments 1 and 2

Experiments 1 and 2 replicated and extended previous results in the literature. The results showed that utilization of a cue declined as the validity of another cue increased and that utilization increased as salience increased, both within and across dimensions. Experiment 1 showed a new result, that the effect of salience interacted with the effect of other-cue validity, although the robustness of this interaction must be treated cautiously in the present data. The results also indicated that learning is very rapid in the initial blocks, with relatively gradual change thereafter. Finally, the results indicated a wide variation across participants, with many participants not utilizing dimensions of positive validity. With these data at hand, we can now pursue quantitative model fitting and from the best fitting parameter values make predictions by the model for the various other situations outlined in the beginning of this article.

The Model: RASHNL

Principles Implemented by RASHNL

RASHNL is an extension of ALCOVE, which is a connectionist model that learns to associate input values with output categories (Kruschke, 1992, 1993b). Mediating this input–output mapping is a stored representation of

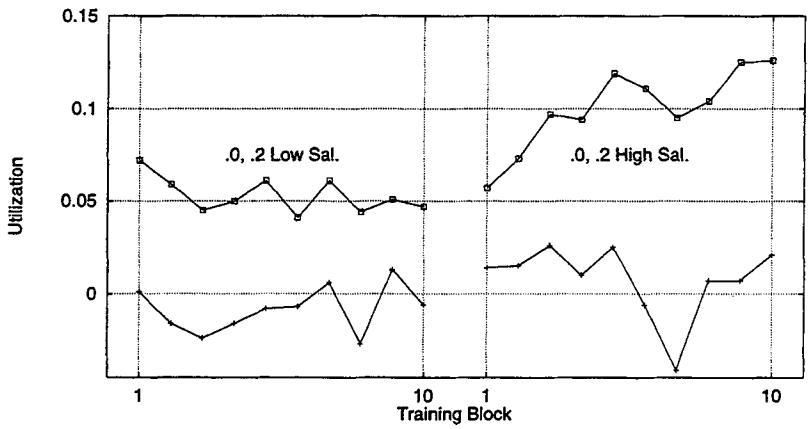

Figure 6. Results of Experiment 2. The abscissa is subdivided into two recurrences of Training Blocks 1–10, corresponding to the two different salience (Sal.) conditions.

previously presented instances of specific combinations of stimulus values. ALCOVE is therefore called an *exemplar based* model. The stimulus values are represented by their coordinates in a separately scaled multidimensional psychological space. A stimulus activates stored exemplars to the extent that the stimulus is similar to the stored exemplars, and similarity is computed by differentially weighting the dimensional differences between each exemplar and the stimulus. The dimensional weights for computing similarity are called *attention strengths,* and they indicate the relevance of the dimension to the categorical distinction being learned. This much of ALCOVE was taken directly from the GCM of Nosofsky (1986), which was generalized from the context model of Medin and Schaffer (1978).

The GCM had no mechanism by which the attention strengths were learned, trial by trial. Instead, they were estimated directly from the data at a certain moment in training, or else the asymptotic attention strengths were predicted to be those values that maximized accuracy. ALCOVE added a learning algorithm both for the attention strengths and for the association weights between exemplars and categories. The learning algorithm was simply gradient descent on the error generated by the model. Thus, learners, on average, were assumed gradually to pay attention to the dimensions that best reduced error, and the same mechanism also drove changes in associative weights. ALCOVE has fit a variety of phenomena in human learning (Choi, McDaniel, & Busemeyer, 1993; Kalish & Kruschke, 1997; Kruschke, 1992, 1993a, 1996b; Nosofsky & Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky et al., 1992; Nosofsky & Palmeri, 1996) but has also been shown to have limitations (e.g., Busemeyer et al., 1993b; Erickson & Kruschke, 1998; Kruschke, 1996a; Lewandowsky, 1995; Macho, 1997).

The original ALCOVE model permitted a capacity limit on attention, such that increasing attention to one dimension entailed decreasing attention to another dimension. This capacity limit was not always imposed, however, because it was not critical for fitting all data sets. One extension of ALCOVE introduced in the present article is a systematic mechanism for limited-capacity attention. In the extended version, each dimensional attention strength is a function of another underlying dimensional variable, namely, the "gain" applied to the dimension. The attention allocated to a dimension is the gain on that dimension, normalized relative to all the other gains. This extension allows the attention strengths to be adjusted smoothly and continuously.

A second extension to ALCOVE, introduced in this article, is the assumption that learners make *large, rapid shifts* of attention on single trials, *before* adjusting associative weights (Kruschke, 1996a). This is contrary to the spirit of the original ALCOVE model, which assumed, but did not require, that learners (or the average of groups of learners) gradually adjusted their attention to dimensions. On the contrary, rapidity of attention shifts is now posited as a critical theoretical principle in accounting for human associative learning. The ADIT [Attention to Distinctive Input] model (Kruschke, 1996a) incorporated this principle in accounting for apparent base-rate neglect and the inverse base-rate effect (Gluck & Bower, 1988b; Medin & Edelson, 1988). The principle can be formally incorporated into ALCOVE because of the new formalism for attentional capacity limitations, mentioned above.

A third extension of the original ALCOVE model is that all learning rates are assumed to be gradually reduced during the course of training, for the types of probabilistic mappings learned in NMCPL. It is assumed that learners adapt to a background level of unavoidable error and learn to discount errors. In lieu of a theory of error discounting, in this article we merely assume that learning rates decline as training progresses. All the various learning rates in the model are assumed to be affected the same way by this "annealing" process. Currently the annealing follows a strict, nonadaptive schedule, but future versions will have to use an adaptive mechanism that adjusts the learning rates or the error signals in response to environmental contingencies.

Figure 7 illustrates the architecture of RASHNL. Each stimulus dimension activates an input node, shown at the lower left of Figure 7. The exemplar nodes are then activated to the extent that they are similar to the input. The similarity is computed with attentionally weighted dimensions, and the attention values are delivered from the attentional network shown at the right of Figure 7. The exemplar node activation propagates up to the output nodes, which represent the various categorical response options. The attentional network on the right of the diagram indicates that attention is a (normalized) function of underlying gain on each dimension and that the dimensional gains are themselves learned associations from context, or bias. The diagram, which is inherently static, cannot indicate the other two extensions of ALCOVE, namely, the rapidity of attention shifts and the annealing of learning rates.

No previous model has simultaneously implemented (a) rapidly shifting, capacity-limited attention, with (b) graded-similarity exemplar representation, and (c) annealing of learning rates. These three principles enable RASHNL to account for many "irrational" phenomena that cannot be addressed by previous models.

Formal Description of RASHNL

Activation propagation. Stimulus dimensions are assumed to be continuous and interval scaled (a case of which

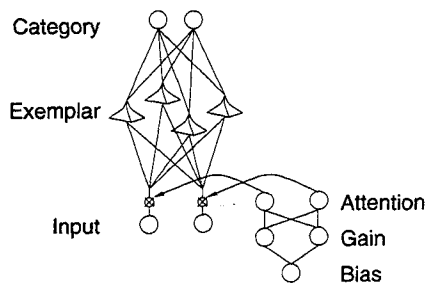

Figure 7. Architecture of RASHNL (Rapid Attention Shifts 'N' Learning). This diagram shows the attention normalization mechanism but does not indicate that attention shifts rapidly on individual trials or that learning and shift rates are annealed.

is binary valued), and each dimension is encoded by a separate input node such that if ψ_i is the psychological scale value of the stimulus on dimension i, then the activation of input node i is

$$a_i^{\text{in}} = \psi_i , \tag{1}$$

where the superscript "in" denotes that this is an input node. We arbitrarily set the scale values of our binary-valued cues to 1 and 2. The behavior of the model depends only on the difference between the scale values, not on their absolute magnitude. For example, the model behaves the same if we use scale values of 0 and 1. The relative salience of different cues is modeled by multiplicative factors on each dimension, as described below.

There is one exemplar node established for each distinct training instance. The experiments modeled here involve just four stimuli, and, for simplicity, all four exemplar nodes were included in the model from the onset of training, rather than being recruited trial by trial.

The activation of an exemplar node corresponds to the psychological *similarity* of the current stimulus to the exemplar represented by the node. Similarity drops off exponentially with distance in psychological space, as suggested by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). Each exemplar node is significantly activated by only a relatively localized region of input space; that is, it has a small "receptive field." Formally, the activation value is given by

$$a_j^{\text{ex}} = \exp\left(-c \sum_{\substack{\text{in} \\ \text{i}}} \alpha_i \sigma_i |\psi_{ji} - a_i^{\text{in}}|\right), \tag{2}$$

where the superscript "ex" indicates that this is an exemplar node, where c is a constant called the *specificity* that determines the overall width of the receptive field, where α_i is the *attention strength* on the i^{th} dimension, where σ_i is the *salience* of the i^{th} dimension, and where ψ_{ji} is the scale value of the j^{th} exemplar on the i^{th} dimension.

Increasing the attention strength on a dimension has the effect of stretching that dimension, so that differences along the dimension have a larger influence on categorization. This attentional flexibility is useful for stretching dimensions that are relevant for distinguishing the categories and shrinking dimensions that are irrelevant to the category distinction (Kruschke, 1992, 1993a; Nosofsky, 1986). A dimension's attention strength denotes the extent to which differences on that dimension are relevant for the category problem at hand. The attention strength does not denote the perceptual discriminability of the dimension.

Like dimensional attention, the salience of a dimension is formalized as a multiplicative factor in Equation 2. Unlike rapidly shifting attention, however, salience is supposed to reflect the underlying perceptual discriminability or intensity of the stimuli, and salience is considered to be stable (i.e., unchanging) during the relatively short course of the experiment. Although Equation 2 includes each dimensional salience as a separate factor, the salience is perhaps better construed as inherent in the psychological scale values of the stimuli. That is, if a dimension is highly salient, then the scale values of the stimuli have a large range. We have parameterized salience as a separate multiplicative factor to give it an explicit referent for our parameter fits.

The attention allocated to the dimensions is nonnegative and capacity constrained. This is accomplished formally by defining dimension i's attention, α_i, to be a function of an underlying *gain*, γ_i, as follows:

$$\alpha_i = \exp(\gamma_i) \Big/ \left(\sum_{\substack{\text{in} \\ \text{j}}} \exp(\gamma_j)^P \right)^{1/P} , \tag{3}$$

where P is a normalization constant that reflects the attentional capacity of the learner. When $P = 1$, the dimensional attention strengths must sum to unity, and any increase of attention to one dimension comes at the cost of the same amount of decrease in attention to other dimensions. When $P \to \infty$, there is no trade-off of attention across dimensions, and all dimensions get full attention all the time. For $P > 1$ but $P < \infty$, there is intermediate attentional competition, so that increased attention to one dimension causes some decrease in attention to other dimensions, but not necessarily as much as the increase. When $0 < P < 1$, any increase in attention to a dimension causes more than that amount of decrease to other dimensions.

The network implementation of this normalization function is illustrated at the right of Figure 7 by the layer of fixed connections between gain nodes and attention nodes. The gains are initialized at zero for all dimensions and can learn to become any real value, as explained below. These learned gain values are suggested in Figure 7 by the connection weights from the "bias" node to the gain nodes.

Activation from the exemplar nodes is propagated to category nodes, which correspond to internal representations of categories. There is one category node per category label. The activation of the k^{th} category node is determined by a linear combination of exemplar-node activations:

$$a_k^{\text{cat}} = \sum_{\substack{\text{ex} \\ \text{j}}} w_{kj}^{\text{cat}} a_j^{\text{ex}} , \tag{4}$$

where w_{kj}^{cat} is the association weight to category node k from exemplar node j. The exemplar-to-category association weights are initialized at zero.

Category node activations are mapped to response probabilities using a version of the Luce (1959) choice rule (i.e., the "softmax" rule in engineering):

$$\Pr(K) = \exp(\phi a_K^{\text{cat}}) \Big/ \sum_{\substack{\text{cat} \\ \text{k}}} \exp(\phi a_k^{\text{cat}}) , \tag{5}$$

where ϕ is a scaling constant. In other words, the probability of classifying the given stimulus into category K is determined by the magnitude of category K's activation relative

to the sum of all category activations. The constant, ϕ, determines the decisiveness of the network: A large value of ϕ expresses a highly decisive choice, in that it causes just a small activation advantage for category K to be translated into a large choice preference for category K. A small value of ϕ expresses an indecisive or unconfident network, in that the small ϕ causes even large activation differences to be translated into ambivalent choices.

Attention shifts. The dimensional attention strengths are shifted by gradient descent on sum-squared error, as used in standard backpropagation (Rumelhart, Hinton, & Williams, 1986). Each presentation of a training exemplar is followed by feedback indicating the correct response, just as in the categorization experiments with human participants. The feedback is coded as *teacher* values, t_k, given to each category node. For a given training exemplar and feedback, the *error* generated by the model is defined as

$$E = \frac{1}{2}\sum_{\substack{\text{cat}\\k}} (t_k - a_k^{\text{cat}})^2 , \tag{6}$$

where the teacher values are defined in these simulations as $t_k = 1$ if the stimulus is a member of category k, and $t_k = 0$ if the stimulus is not a member of category k.[2]

On presentation of a training instance, learning progresses in two steps. First, attention is rapidly shifted to reduce error. Second, after attention is shifted, the association weights are adjusted to reduce any remaining error. These two steps are discussed in turn.

Attention is shifted proportionally to the (negative of the) error gradient with respect to the dimensional gains. Evaluating the gradient leads to the following formula for Dimension A's attention shift:

$$\Delta\gamma_A = -\lambda_\gamma \sum_{\substack{\text{in}\\i}} \sum_{\substack{\text{ex}\\j}} \sum_{\substack{\text{cat}\\k}} (t_k - a_k^{\text{cat}}) w_{kj}^{\text{cat}} a_j^{\text{ex}} c \times \sigma_i |\psi_{ji} - a_i^{\text{in}}| (\kappa_{iA}\alpha_A - \alpha_i \alpha_A^P) , \tag{7}$$

where λ_γ is a nonnegative constant of proportionality called the *shift rate* for attention, and $\kappa_{iA} = 1$ if $i = A$ and is zero otherwise (κ_{iA} is sometimes called the *Kronecker delta function* of i and A, but we avoid using the symbol δ to avoid confusion with error terms in the "delta rule" of backpropagation). A derivation of Equation 7 appears in Appendix B.

Psychologically, attention is hypothesized to shift a large extent on a single trial. This large shift cannot be achieved formally with a single large step in the direction of the gradient because attention is a highly nonlinear function of gain; that is, the gradient changes as the attention changes. Therefore, the change specified by Equation 7 is iterated 10 times (an arbitrary number) on each trial, so that the nonlinearity of the function can be approximated with 10 relatively small steps. After each small attention change, the activation is repropagated to the category nodes to generate a new error, and attention is changed a small amount again, for 10 iterations. The result of these 10 small steps constitutes the single large shift.

Learning of associations. After the attention is shifted, the association weights are adjusted, also by gradient descent on error, which yields the following formula for associative weight changes:

$$\Delta w_{kj}^{\text{cat}} = \lambda_w (t_k - a_k^{\text{cat}}) a_j^{\text{ex}} , \tag{8}$$

where λ_w is a nonnegative constant of proportionality called the *learning rate.*

The associative weights from the bias node to the gain nodes are also adjusted via gradient descent on error, where *error* is defined as the difference between the shifted value and the initial, preshift value. That is, the shifted value acts as the teacher, or target, for the gain nodes. To reduce the number of free parameters, the learning rate for gain biases was arbitrarily set to 1.0. In principle, the learning rate for gain biases should be less than 1.0, so that the model's learned attention to dimensions will perseverate in situations when dimensional relevance shifts, just as humans appear to do (e.g., Kruschke, 1996b). This is not critical in the current application, however, because the annealing of learning rates (described in detail below) quickly makes the effective learning rate much less than 1.0.

Annealing schedule. The shift rate for gains, and the learning rates for category association weights and for gain biases, were all gradually reduced across trials of training, to reflect the fact that human learners appear also to "turn off" their learning, or discount errors, in these probabilistic situations. In the current model, this reduction was achieved by a commonly used "annealing" schedule, whereby the shift and learning rates were multiplied by a factor, $r(t)$, that decreased with training trials, t, as follows:

$$r(t) = 1/(1 + \rho t) , \tag{9}$$

where ρ is a freely estimated scheduling constant. This formula for annealing is called a "search then converge" schedule by Darken and Moody (1991), because the learning rate stays relatively high until trial $1/\rho$, and then decreases with training trials, as displayed in Figure 8. Annealing schedules are discussed further at the end of the article.

Summary of free parameters. In fitting RASHNL to human learning data of Experiments 1 and 2, there are the following nine free parameters: the probability mapping constant ϕ in Equation 5; the exemplar-to-category association weight learning rate λ_w in Equation 8; the exemplar specificity c in Equation 2; the attention strength shift rate λ_α in Equation 7; the attentional capacity P in Equation 3; the annealing schedule constant ρ in Equation 9; and three relative salience values, σ_i in Equation 2, for the three different magnitudes of rectangle height variation. For

[2] The teacher values used in these simulations were "strict," not "humble" like the teacher values used in the original ALCOVE model (Kruschke, 1992). Strict teachers were used merely for simplicity, because use of humble teachers made no difference in the fits.

simplicity, the relative salience of the line segment position was arbitrarily fixed at 1.0, regardless of the different rectangle heights. This simplification might not have been entirely appropriate, however, because the salience of the line segment might have depended on the magnitude of rectangle height variation. Mellers and Biagini (1994) reviewed several cases in which smaller contrasts on one dimension produce greater psychological effects for the other dimension.

Fit of RASHNL to Effects of Competing-Cue Validity and Salience in Experiments 1 and 2

RASHNL was fit simultaneously to all 140 data points of the learning curves from both Experiments 1 and 2, shown in Figures 4 and 6. The best fit yielded a root-mean-square deviation (RMSD) of .0207 with parameter values as follows: $\phi = 3.55$, $\lambda_w = 0.152$, $\lambda_\alpha = 48.6$, $\rho = 0.337$, $P = 2.95$, $c = 10.3$, and the saliences of the rectangle variations relative to the line segment variation equal to 0.783, 0.883, and 0.894, corresponding in rank to the physical differences. (The value of the attention shift rate, $\lambda_\alpha = 48.6$, should not be appraised as untenably large. The error signal that drives the attention shift is highly attenuated by being backpropagated through the nonlinear exemplar nodes and the normalization function, so, unlike associative weight learning rates such as λ_w for linear nodes, the largest sensible learning rate for λ_α is not capped at 1.0.) The model accounts for 94.9% of the variance in the 140 data points, and a plot of predicted utilizations as a function of actual utilizations appears in Figure 9.

Figures 10 and 11 show the best fitting utilizations by RASHNL, for Experiments 1 and 2 (compare with the empirical results in Figures 4 and 6). The model shows the major trends in the empirical data: Utilization rises rapidly in the first three blocks and rises only slowly thereafter; utilization declines as the validity of the competing dimen-

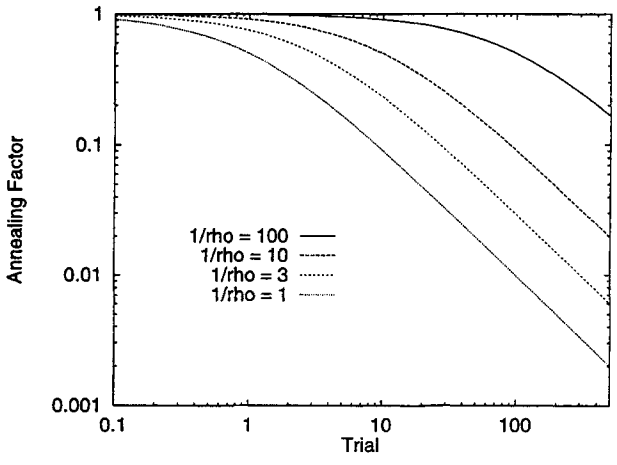

Figure 8. Plots of the annealing factor as a function of the learning trial, corresponding to Equation 9. The annealing factor remains fairly high until trial 1/rho and then approaches a linear decline (for log–log scales).

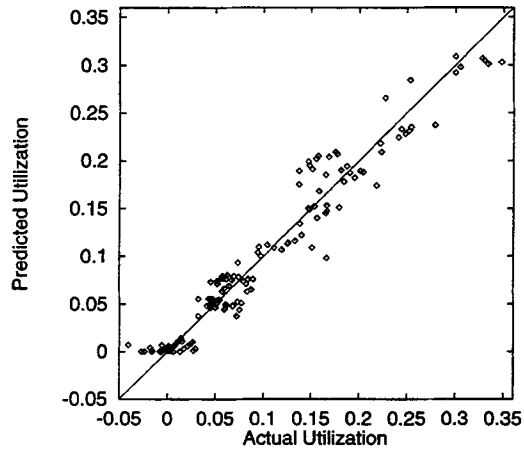

Figure 9. Best fitting (predicted) utilizations plotted as a function of actual utilizations, for the 140 data points. The model accounts for 94.9% of the variance in the data.

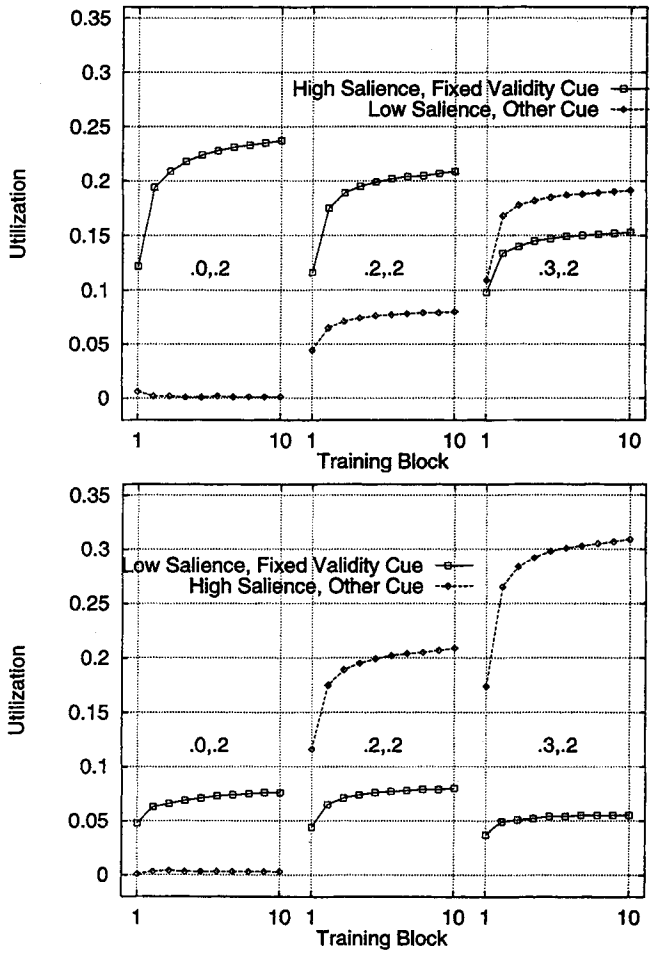

Figure 10. Best fit of RASHNL (Rapid Attention Shifts 'N' Learning) to results of Experiment 1 (when also fit simultaneously to results of Experiment 2). Compare with Figure 4.

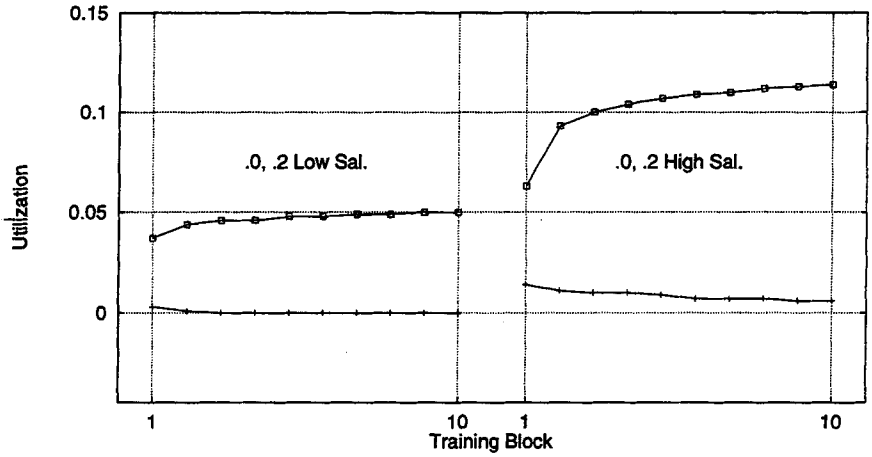

Figure 11. Best fit of RASHNL (Rapid Attention Shifts 'N' Learning) to results of Experiment 2 (when also fit simultaneously to results of Experiment 1). Compare with Figure 6. Sal. = salience.

sion increases, and utilization increases with increasing salience. The model also shows an interaction between salience and other-cue validity (observed in Experiment 1), wherein the effect of salience of the .2-validity cue is larger when the other cue has 0 validity than when the other cue has .3 validity.

Quantitative deviations from the data are unsystematic except for Block 1, where the model tends not to utilize the cues as much as people do, and except for the .2, .2 condition of Experiment 1, where the model shows a larger effect of salience than is seen in the data. The high performance by people in Block 1 is probably attributable in part to their use of STM for previous trials, a mechanism not implemented in the model. Previous work in our lab (Kruschke & Bradley, 1995; Kruschke & Erickson, 1995) showed that STM could account for deviations of delta-rule learning curves from early trials of human learning. It is plausible that incorporating an STM mechanism in RASHNL would address the early-trial deviations seen here. The model's oversensitivity to salience in the .2, .2 condition might be ameliorated if an adaptive annealing schedule were implemented, as opposed to the fixed annealing schedule used here. An adaptive annealing schedule would, presumably, create a larger annealing rate for the .0, .2 and .2, .2 conditions than for the .2, .3 conditions, because the former are more probabilistic than the latter. This relative decrease in the annealing rate for the .2, .3 condition would cause both learning curves in the .2, .3 condition to rise relative to the other curves, and compensatory adjustments in the other parameter values could create a better fit. These speculations await future testing with specific implementations. In the meantime, the current fit is informative, and the model is shown in subsequent sections to predict many other phenomena in multiple-cue probability learning.

Figure 12 shows the predicted utilizations by individual participants in Experiment 1 (compare with Figure 5). The model does not show as much variance as human participants show; in particular, the model does not show the high (i.e., nearly .5) utilizations observed in some human participants. But this limitation is caused by the fixed scaling constant in the choice function (Equation 5): The value of ϕ puts an upper bound on the possible utilization exhibited by the model, because output node activations are taught to be in the limited range [0, 1].

This problem could be readily addressed by sampling ϕ for each simulated participant from a random distribution, so that ϕ is large for some participants and small for others. Increasing or decreasing ϕ, while leaving all else constant, merely increases or decreases the utilization and does not affect any aspects of learning. With regard to Figure 12, this sort of randomization of the ϕ values would produce random, approximately *radial* shifts of the points, so that some points would become further from the origin and some closer to the origin. The psychological interpretation of this randomization is straightforward: The value of ϕ reflects the participant's confidence or decisiveness in converting relative activations of categories to overt choice preferences. A large ϕ indicates that the participant is willing to translate a small relative advantage of one category into a strong choice preference for that category. It is plausible that not all participants approach this task with equal "confidence" in their responses, so that different participants should be modeled with different ϕ values. At this time, however, we have no theory to motivate a particular distribution of ϕ over individuals, and this issue is an avenue for future research. Other sources of random variation are also plausible: Across participants, there may be individual differences in the perceived saliences of the cues. Within participants, there may be trial-to-trial fluctuations in motivation and overall attention, which could be formalized as noise in the learning rates, attention capacity, or specificity.

From another perspective, however, RASHNL shows a surprisingly large degree of variance, compared with what one might expect from gradient-descent learning models generally. Many models that adjust variables by gradient descent converge to a common asymptotic value regardless of the random training sequence (Busemeyer et al., 1993b). Indeed, annealing schedules are typically designed to encour-

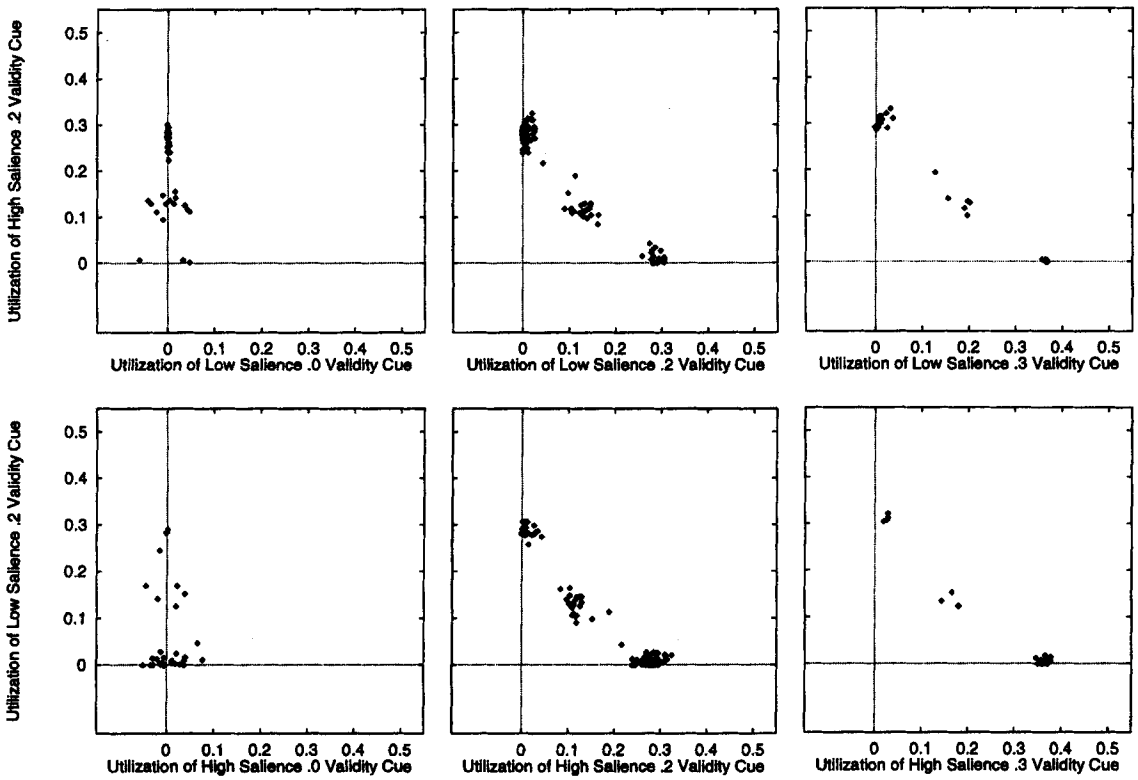

Figure 12. Utilizations of the two cues by individual simulated participants in RASHNL (Rapid Attention Shifts 'N' Learning) for Experiment 1. Compare with Figure 5.

age rapid convergence to asymptote despite probabilistic training patterns (e.g., Darken & Moody, 1991). The utilizations plotted in Figure 12 are averaged over the last 280 trials of training, so they represent fairly stable utilizations, not transitory states. Particularly remarkable is the fact that many simulated participants in RASHNL do not utilize dimensions of nonzero validity, even after hundreds of training trials, just as seen in human participants. This selective nonutilization is caused in the model (and presumably in humans) by rapidly shifting attention: On individual trials, one dimension proves to conflict with recently learned associations, and so attention is shifted strongly away from the conflicting dimension to the consistent dimension, where it tends to stay.

The importance of rapidly shifting attention is highlighted by examining the best fit of the model when this mechanism is removed. When the attention shift rate is fixed at zero, the best fitting RMSD worsens noticeably to .0298 (from .0206), with parameter values of $\phi = 2.80$, $\lambda_w = 0.430$, $\rho = 0.586$, $P = 3.48$, $c = 11.9$, and saliences of 0.0665, 0.137, and 0.137 for the three different rectangle height variations (relative to the line segment salience fixed at 1.0). Without attention shifts, the model shows very little decline in utilization of the high-salience .2-validity cue in Experiment 1, with last-block utilizations of .214, .205, and .195 in the .0,.2, .2,.2, and .3,.2 conditions, respectively. The no-attention model also shows an ordinal violation of the data: For Experiment 1, when the low-salience rectangle height has validity .3 and the high-salience line position has validity .2, the no-attention model predicts that the .3-validity cue is utilized .050 *less* than the .2-validity cue, contrary to the data, in which the the .3-validity cue is used .023 *more* than the .2-validity cue (see Figure 4). The difference in the data is not significant, $t(31) = 0.59$, $p = .56$, but the difference in the predictions is highly significant, $t(31) = 6.15$, $p < .0001$.

With attentional shifting fixed at zero, the individual participant utilizations are also too tightly clustered, and the model cannot learn to selectively ignore a cue. Figure 13 shows the utilizations by individual simulated participants in the .2, .2 condition of Experiment 1, when the attention shift rate is fixed at zero. Notice that none of the individual utilizations for either dimension is at zero, for all 112 simulated participants.

Annealing also plays a critical role in the model's behavior. When the annealing rate is fixed at zero, the best fitting RMSD is poor, at .0449. Without annealing, the model can still show some of the effects of salience and other-cue validity because of rapidly shifting attention and the nonlinear choice function, but the quantitative fit is very poor because the model tends to converge to a common asymptote in all conditions. Annealing is crucial for fitting the data, because it causes the influence of early learning to be nearly "frozen" into the network.

RASHNL exhibits cue competition because of its rapid shifts of limited capacity attention. When one dimension is

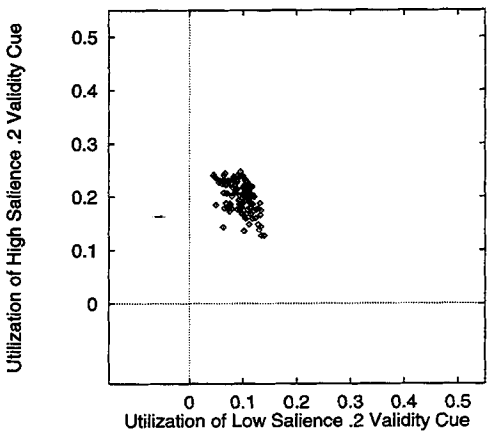

Figure 13. Utilizations of the two cues by individual simulated participants in RASHNL (Rapid Attention Shifts 'N' Learning) *with no attention shifts* for the .2, .2 condition of Experiment 1. Compare with the upper middle panel of Figure 12.

more valid than another, attention tends to shift toward the more valid dimension at the cost of reduced attention to the less valid dimension. The more that attention is given to a dimension, the more that the dimension will be learned about. These attentional shifts have a strong influence on early learning and could be compensated for by later learning, were it not for the fact that the early influences are virtually frozen into the network by annealing of the learning rates.

RASHNL utilizes higher salience dimensions more than lower salience dimensions because of the way the two different values of the cues are represented by the similarity-based exemplar nodes. Recall that the exemplar nodes have overlapping receptive fields, so that, for example, the short rectangle stimulus will partially activate the tall rectangle exemplar node. When a dimension has low salience, the exemplar nodes representing the two cue values are greatly overlapping, so that there is relatively little difference in the exemplar node activations for the two different cue values. This poor discrimination of the cue values causes slow learning about the dimension. Because early learning is frozen into the network, slow learning implies little learning.

Edgell et al. (1996) and Edgell et al. (1992) suggested that the effect of salience on utilization was caused by greater confusion in STM. Lower salience cues are confused more in STM, producing a lower effective validity of the cue. Of course, lower validity cues are utilized less. In some respects, RASHNL implements just such a mechanism. If the exemplar node activations are conceived of as STM activations, then lower salience cues are indeed confused more, insofar as the exemplar node activations are less discriminating between cue values.

Predictions of RASHNL for Other Situations

In the next few sections of the article, we present predictions of RASHNL for several different situations. The goal of these sections is to show RASHNL's *qualitative* predictions of various effects and interactions, *using the fixed parameter values* that best fit the data from Experiments 1 and 2. The aim of the predictions is not quantitative accuracy, because other experiments that are reported in the literature and that investigated these other situations, used somewhat different stimuli, apparatus, instructions, and so on. The exact levels of utilization in any particular experimental situation depend critically on the saliences of the particular stimulus dimensions, the perceptual separability of the dimensions, the procedural details of training, and so forth. We assume, however, that the other experiments were similar enough to ours that the parameter values that best fit our experiments should produce at least qualitatively correct predictions for the other situations. Using these fixed parameter values, we show that RASHNL does indeed make qualitatively correct predictions for a variety of effects and interactions reported during two and a half decades of research by Edgell and colleagues. We also find that RASHNL makes two novel predictions, regarding an interaction of cue salience with additional irrelevant cues and regarding an effect of irrelevant-cue salience. These novel predictions are subsequently confirmed in Experiments 3 and 4.

Dimensional Utilization With Additional Irrelevant Cues

Castellan (1973) reported that utilization of a partially valid cue decreased as additional irrelevant cues were added to the stimulus display. For example, consider the .0, .2 condition of Experiment 1 in which the rectangle height had a validity of .2 and the line segment position had a validity of zero. Castellan's (1973) results suggest that people's utilization of the rectangle height would be larger if the irrelevant line segment was absent from the display, and people's utilization of the rectangle height would be smaller if another irrelevant cue were added to the display.

Castellan (1973) also reported that the influence of irrelevant cues interacted with the validity of the primary cue. Thus, when the primary cue had very high or very low validity, the additional irrelevant cues had relatively little effect on utilization. The detrimental effect of additional irrelevant cues was felt most strongly when the primary cue had an intermediate validity.

To test the predictions of RASHNL, the model was trained in conditions such that one dimension had nonzero validity (either .1, .2, .3, or .4) and was presented either by itself or with other dimensions of zero validity. For each condition, 200 simulated participants (random sequences) of 400 trials were generated in blocks of 80 trials that realized the validities perfectly. The same parameter values that best fit the results of Experiments 1 and 2 were used in these simulations, with the salience of each dimension set to 1.0.

Figure 14 shows that RASHNL predicts the empirically observed effect and interaction. Figure 14 plots the model's mean utilization of the nonzero validity dimension in the last 40 trials. The model's utilization of the relevant dimension declines when an irrelevant dimension is included, and the decline is greatest for middling validities. Not shown in the

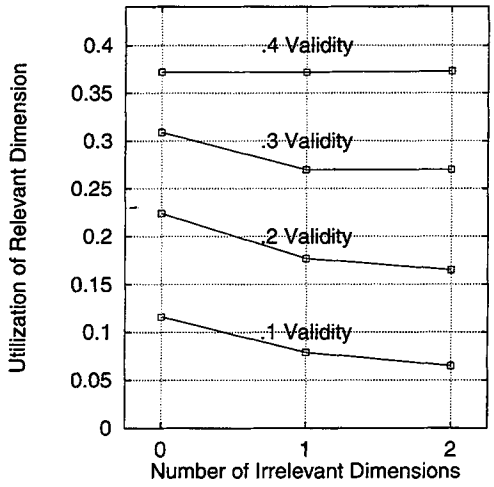

Figure 14. Utilization by RASHNL (Rapid Attention Shifts 'N' Learning) of a valid dimension as a function of its validity and of the number of irrelevant (zero-validity) additional cues.

figure is the fact that at near-zero validities the model shows near-zero utilizations, regardless of the number of dimensions.

RASHNL shows the decline in utilization with the addition of irrelevant cues because of competition for attention. When additional dimensions are added, they receive some attention at the cost of reduced attention to the relevant dimension. The model does learn to allocate more attention to the relevant dimension than to the irrelevant dimensions but not very strongly, because attending to the relevant dimension occasionally leads to error and because the learning and shift rates are annealed. This effect of additional dimensions is reduced at higher validities because the higher validities more rapidly drive attention to the relevant dimension and more rapidly build larger associative weights to the categories, before annealing "freezes" the learning.

The absolute levels of utilization by the model are a bit low for the high-validity conditions, compared with human utilizations, but this could be corrected by an adaptive annealing rate, as opposed to the fixed annealing rate implemented in the displayed simulations. For higher validity cues, the mapping is less probabilistic, hence the annealing rate should be less. This lesser annealing allows greater utilization of the cues. Additional simulations verified that when the annealing rate is reduced proportionally to the cue validity, the trends of Figure 14 persist, but the absolute utilizations increase for higher validity cues.

Two Novel Predictions: Effects of Irrelevant-Cue Salience

A novel prediction of RASHNL is that the effect of adding an irrelevant cue should interact with the saliences of the cues. In other words, the effect of cue salience should depend on the number of irrelevant cues. Consider, for example, the effect of salience observed in Experiment 2. The large variation, higher salience rectangle had mean utilization of .110, and the small variation, lower salience rectangle had mean utilization of .051. These results occurred in the presence of an irrelevant, zero-validity line segment variation. If the irrelevant line segment is removed, RASHNL makes the following predictions: First, as shown above, utilizations of both the higher and lower salience rectangles should increase. Second, and this is the novel prediction, the effect of salience should be smaller; that is, the difference in utilizations between the high- and low-salience rectangles should be less. Later in the article, after the collection of effects accumulated by Edgell and colleagues has been addressed, Experiment 3 verifies this prediction, and RASHNL is shown to fit the data well.

A second novel prediction is that the salience of an irrelevant cue will differentially affect the utilization of a relevant cue. Specifically, when the irrelevant cue is of relatively high salience, then the relevant cue will be utilized less than when the irrelevant cue is of relatively low salience. RASHNL predicts fairly large effects of irrelevant-cue salience, but Edgell and colleagues (Edgell et al., 1992, 1996) have found only nonsignificant trends in this direction and have concluded that no such effect exists. Experiment 4 of this article tests and confirms this RASHNL prediction, with data well fit by the model.

Base-Rate Utilization With Additional Irrelevant Cues

Outcome base rates fall in the range of 0 to 1, whereas dimensional validities have values in the range of $-.5$ to .5. This is because the outcome base rate expresses the mean probability of category K, which can only be between 0 and 1, whereas dimensional validities express deviations from this mean. For example, an outcome base rate of 0.7 indicates that on average across all trials, without regard to particular cue values, category K occurs 70% of the time.

Edgell and Hennessey (1980) reported that when the outcome base rate was 0.7, utilization of the base rate declined as the number of irrelevant cues increased from 1 to 3. This result is analogous to the results of Castellan (1973), summarized earlier, which showed that utilization of a relevant dimension decreased as irrelevant dimensions were added. Edgell and Hennessey (1980) also found that the decline in base-rate utilization was smaller for more extreme base rates, analogous to the interaction found by Castellan (1973) wherein the decline of dimensional utilization was smaller for more extreme dimensional validities.

Like the dilution of dimensional utilization, the dilution of base-rate utilization can be captured within the framework of RASHNL. Just as the dilution of dimensional utilization was explained by reduced allocation of attention to the relevant dimension, the dilution of base-rate utilization will be explained by reduced allocation of attention to a cue that conveys base-rate information.

What cue conveys base-rate information? Consider classical studies of base-rate learning (e.g., Estes & Straughan, 1954) in which learners saw nothing at the onset of a trial except a response prompt, which indicated that it was time to guess which one of two outcomes would subsequently

occur. In these studies, the participants were implicitly instructed that when there was no response prompt, neither outcome occurred, but participants learned that when there was a response prompt, Outcome 1 occurred with some probability and Outcome 2 occurred with the complementary probability. Thus, when the base rates of the categories were unequal, the mere occurrence of the response prompt conveyed information about the probable outcome; that is, the response prompt was a cue of nonzero validity.

We are presently interested in the utilization of response-prompt information when irrelevant cues are added to the response prompt. With this interpretation of the response prompt as a cue for base rates, the results of Castellan (1973) imply that utilization of response-prompt information should be diluted when irrelevant cues are added. Moreover, their results imply that there should be an interaction such that the dilution should be weak for highly valid response prompts, but the dilution should be stronger for middling validity response prompts. The results of Edgell and Hennessey (1980) verify both of these implications.

With the inclusion of a dimension to represent the response prompt, the explanation of the dilution of base-rate utilization is the same as the explanation for the dilution of dimensional utilization (recall Figure 14): There is competition of attentional allocation to the various dimensions, including the response-prompt dimension, and so utilization of the response prompt declines as more dimensions are added.

This notion of the response prompt acting as a cue for base-rate information is an enhancement of the ADIT model (Kruschke, 1996a), which was the first model to implement rapid error-driven shifts of attention. In ADIT, the dilution of base-rate utilization was caused by a separate response mechanism. This separate mechanism is jettisoned by RASHNL, which takes a unified approach to dimensional and base-rate utilization.

The response prompt was not included in the input representation for the simulations reported above because (a) it is needed only for cases when base rates are unequal, (b) it would demand additional free parameters such as the salience of the response prompt relative to the other cues, and (c) it would require specification of several ad hoc assumptions regarding a variety of properties such as the initial weights between values of the response prompt and the output nodes, or perhaps the learning of no response during intertrial intervals. These complications do not, however, change the basic qualitative predictions of the model, which is based on limited capacity attention allocated to the presented dimensions.

Predictions of RASHNL for Configural Utilization

Recall from the beginning of our article that the configuration of cues is informative when the probability of an outcome, given a cue combination, cannot be perfectly predicted by a linear combination of the marginal probabilities of the cues. In other words, the outcome probability is correlated with the logical XOR of the cue values. In this section we explore the predictions of RASHNL for utilization of configural information.

Configural Validity Is Utilized Less Than Dimensional Validity

Edgell (1978) and Edgell and Castellan (1973) found that configural information was utilized less than dimensional information, even when the two had equal validities. RASHNL was tested for its ability to predict this behavior. As in the human experiments, the model was trained on stimuli with two cues. The parameter values were the same as the best fit for Experiments 1 and 2, with both dimensional saliences set to 1.0. Predictions of the model were determined by the mean utilization of 200 randomly generated sequences (participants), in the last 40 trials of 400 training trials.

Table 1 shows the predicted utilizations for various combinations of dimensional and configural validities. The first and second rows of the table indicate utilizations when only a dimension has validity or only the configuration has validity, and it can be seen that configural information is utilized much less than dimensional information (utilizations of .127 vs. .169). When dimensional validity and configural validity are simultaneously present, as reported in the third row of the table, the configural utilization is again less than the dimensional utilization (.102 vs. .165). Thus, the model makes qualitative predictions that match human results.

The model's lesser utilization of configural information is understandable in terms of limited-capacity attention: Configural information requires partial attention to both dimensions, but dimensional information requires attention to only one dimension. When attention is distributed over dimensions, the overlapping receptive fields are closer together, causing more interference and slower learning. When dimensional and configural information are simultaneously present, the informative dimension attracts attention, at the cost of reducing attention to the uninformative dimension, and hence additional interference with configural information.

Dimensional Salience and Configural Utilization

Edgell (1993, pp. 52–53) reported that when two dimensions have very different saliences, then the use of the low-salience dimension can be as low as the utilization of

Table 1
Predicted Utilization of Dimensional and Configural Information by RASHNL

Validity			Utilization		
Cue A	Cue B	Config.	Cue A	Cue B	Config.
.2	0	0	.169	0	0
0	0	.2	0	0	.127
.2	0	.2	.165	0	.102

Note. RASHNL = Rapid Attention Shifts 'N' Learning; Config. = configural information.

configural information. In an experiment conducted with Pak Ng, the researchers compared three conditions: Utilization of a high-salience cue of .2 validity, utilization of a low-salience cue with .2 validity, and utilization of the .2-validity configuration of the high- and low-salience cues. They found that the low-salience dimension was utilized "at about the same level as the configural information."

RASHNL qualitatively reproduces this result. Using the same parameter values as the best fit to Experiments 1 and 2, we let one dimension have a salience of 1.1 and the other dimension have a salience of 0.9. Again we ran each condition with 200 randomly generated sequences of 400 trials, and we computed the mean utilization for the last 40 trials. When the high-salience dimension had a validity of .2, its mean utilization by RASHNL was .267. When the low-salience dimension had a validity of .2, its mean utilization was .071. Finally, when the configuration had a validity of .2, its mean utilization was .068, essentially the same as the utilization of the low-salience dimension.

Interaction of Component and Configural Utilization

A number of studies by Edgell and collaborators have investigated the interaction of dimensional and configural validity when the valid dimension either is or is not one of the dimensions in the valid configuration. For example, Edgell (1980, Experiment 2, Condition 3, and Experiment 3, Condition 3) examined utilization of information in three-dimensional stimuli. In one condition, the first dimension had validity .2, and the configuration of the first and second dimension had validity .3. In another condition, the first dimension again had validity .2, but the configuration of the second and third dimensions had validity .3. In other words, in the first condition the relevant configuration included the relevant dimension, but in the second condition the relevant configuration involved different dimensions. Edgell found that configural utilization dropped significantly from the first condition to the second; that is, when the configural information no longer overlapped with the dimensional information, the configural utilization dropped.

Although Edgell (1980) did not discuss salience, it is important to note that the first dimension in these studies was the orientation of stripes. Subsequent work by Edgell et al. (1992) showed that stripe orientation is much more salient than the other dimensions. To reflect this salience difference, our simulations set the salience of the first dimension to 1.2 and the saliences of the second and third dimensions to 0.9.

RASHNL predicts the following results: When the first dimension has validity .2 and the configuration of first and second dimensions has validity .3, then the utilization of the first dimension is .281 and the utilization of the configuration is .090. When the first dimension has validity .2 and the configuration of second and third dimensions has validity .3, then the utilization of the first dimension is .222 and the utilization of the configuration is .034. That is, the utilization of the configural information decreases (from .090 to .034) under these conditions, which is qualitatively the same result seen in human participants by Edgell (1980).

Edgell and Roe (1995) examined interactions of component and configural utilization when the cue saliences were approximately equal. In their Experiments 1 and 2, they measured utilization of configural information when it occurred by itself, when it occurred with a third dimension also having validity, and when it occurred with one of its own dimensions having validity. As found in previous studies, the magnitude of configural utilization was much less than the magnitude of dimensional utilization. The level of configural utilization depended, however, on whether or not the valid dimension was a component of the valid configuration. Configural utilization was higher when the valid dimension was a component of the configuration. Dimensional utilization was not significantly different when the valid dimension was part of the configuration or not. Overall, then, there was competition between dimensional and configural utilization, with the effects of competition manifested predominantly in the configural utilization.

When RASHNL is applied to this situation, the model also shows competition between dimensional and configural utilization, as should be expected from its limited-capacity attention. However, in the model the competition is manifested in the dimensional utilization, such that the configural utilizations remain at approximately the same low level, while the dimensional utilization is larger when the valid dimension is one of the configural dimensions than when it is a separate dimension. This prediction, however, depends on using the parameter values that best fit our data. If the model is instead fit directly to the data from Experiments 1 and 2 of Edgell and Roe (1995), an extremely accurate fit can be obtained with different parameter values. What appears to be most important for fitting their data is an attentional capacity value that is less than unity; that is, in Equation 3, $P < 1$. When $P < 1$, full attention can be given to any single dimension, but if attention must be distributed over N dimensions, each dimension gets an attentional strength that is *less* than $1/N$. Because of this penalty for attending to multiple dimensions, configural utilization suffers more in the attentional competition than does dimensional utilization. As of yet we have no theory to explain or predict the value of the attentional capacity P, but presumably its value could be influenced by a variety of factors such as the motivation of the learner, the perceptual integrality of the cues, the ease of processing each cue individually, the number of cues, the instructions given to the participants, and so on. In any case, the theoretical principles, formalized by the model, remain tenable.

Predictions of RASHNL for Delayed Introduction of Cue Validity

Edgell (1983, Experiment 1) showed that when configural information was introduced into training after a delay during which dimensional information was already present, the configural information was utilized much less than when it was introduced at the beginning of training. Edgell compared one condition, in which the first dimension had a validity of .2 and the configuration of the first and second dimensions had validity of .2 throughout 400 trials of training, with another condition, in which the configural

validity of .2 was introduced only after 120 trials with zero configural validity. His results showed that the utilization of the configural information was much less in the delayed group than in the from-the-start group, even at the end of 400 trials. Edgell (1983, Experiment 2) and Edgell and Morrissey (1987, Experiment 1), showed that when the configural information was introduced at various trials (20, 40, 80, 120, and 200), utilization of the configural information rose to roughly the same level in all the delayed groups but never to as high a level as the from-the-start group.

RASHNL qualitatively reproduces these results. Using the same parameter values as before, RASHNL learns configural information much better when it is introduced from the start than when it is introduced later. Table 2 shows the utilizations predicted by RASHNL as a function of the number of trials in the delay (for 200 simulated participants, in the last block of 40 trials out of 400 training trials). RASHNL shows a large reduction in utilization after just a 20-trial delay, and the reduction after that is relatively small and gradual.

In RASHNL, this drop in learning is caused by annealing. The "search then converge" annealing schedule causes the model to be sensitive to information that is present during the early "search" phase, but the schedule causes the model to be relatively insensitive to information present only during the later "converge" phase. The learning rate in this annealing schedule never drops all the way to zero, but it does get close to zero quickly after the first few trials (see Figure 8).

This fixed annealing schedule is, as mentioned before, merely a proxy for a truly adaptive learning rate to be explored in future research. A fixed annealing schedule might be appropriate for a stationary environment, in which the probabilities of contingencies never change. In contrast, for a nonstationary environment such as delayed introduction of information, the learning rates should adaptively respond to changes in contingencies. Presumably such an adaptive mechanism would track the consistency of errors being made and the extent to which learning can reduce the errors. If errors cannot be reduced with learning, then they should be discounted. On the other hand, if errors are effectively reduced with learning, then the learning rates should increase. Possible formalisms for such adaptive learning rates are discussed later.

Edgell and Morrissey (1987, Experiment 2) showed that when *dimensional* information was introduced after a delay, it was learned more slowly than when the same information was introduced from the start. By the end of 400 training trials, however, utilization rose to almost the same level as dimensional information introduced from the start. In their study, the relevant dimension was more salient than the irrelevant dimension, and the configuration of dimensions had a constant validity of .2. RASHNL reproduces the slower learning of delayed dimensional information but does not utilize the dimensional information to the same extent at the end of 400 trials of training. When the dimensional information is introduced from the start, the utilization in the last block is .254. When the dimensional information is introduced after 40 trials, the utilization in the last block is .150. (These results are for saliences of 1.1 and 0.9 for the relevant and irrelevant dimensions, respectively.) This partial success and partial failure of RASHNL is again attributable to the fixed annealing schedule. The annealing causes learning rates to decline, so that information introduced after a delay is learned more slowly than information introduced at the start. But the fixed annealing schedule causes the learning rates to remain small, even after the contingencies have changed, so that new information is learned too slowly. A more complete model would use adaptive learning rates, rather than a fixed annealing schedule. This is described at greater length in the General Discussion.

Table 2
Predicted Utilization of Configural Information Introduced After a Delay

Delay (trials)	Utilization
0	.100
20	.057
40	.043
80	.034
120	.023
200	.016

On the other hand, Edgell and Morrissey (1987, Experiment 3) showed that when dimensional information was introduced after a delay, it was *not* necessarily utilized as much as when it was introduced from the start, even at the end of 400 trials. In their Experiment 3, rather than the configuration having constant validity as in their Experiment 2, the low-salience dimension had constant validity of .2, while the high-salience dimension had validity changing from .0 to .2 after a delay. RASHNL does qualitatively reproduce their results: When introduced from the start, the high-salience dimension is utilized to a level of .244 at the end of 400 trials, whereas after an 80-trial delay, it is utilized only to a level of .120.

In summary, these simulations of delayed introduction of information have both positive and negative implications for the model. On the positive side, the simulation results demonstrate that the annealing mechanism is important to the model for addressing the slower learning of information introduced after a delay. On the negative side, the simulation results demonstrate the (perhaps obvious) point that a fixed annealing schedule is inappropriate for a nonstationary environment. Adaptive annealing schedules for nonstationary regimes are an unsettled topic in the stochastic optimization literature and have been only rarely applied in experimental psychology. This is potentially a rich domain for future research and is discussed more below.

Experiment 3: Interaction of Additional Irrelevant Dimensions and Salience

As discussed in a previous section, a prediction of RASHNL is that the deleterious effect of adding an irrelevant dimension should be stronger when the relevant

dimension is of low salience than when it is of high salience. We now report results of an experiment that confirmed this prediction. In this experiment, an irrelevant cue of intermediate salience was added to a relevant cue that had either high or low salience. When the relevant cue had high salience, then the addition of a moderate-salience irrelevant cue was predicted to have only a small deleterious effect on the utilization of the relevant cue. When the relevant cue was of low salience, however, then the addition of a moderate-salience irrelevant cue was predicted to have a large effect on the utilization of the relevant cue.

Method

Participants. A total of 193 students from introductory psychology courses at Indiana University volunteered for partial course credit.

Design. For every participant, there was one relevant dimension of .2 validity. The relevant dimension could be of high or low salience. The relevant dimension could appear by itself or could be accompanied by an additional, irrelevant (.0 validity) dimension of intermediate salience. Thus, the experiment comprised a 2 × 2 between-subjects factorial design, crossing relevant-dimension salience (high or low) with number of irrelevant dimensions (zero or one). As in Experiments 1 and 2, training consisted of 10 blocks of 40 trials. Each block exactly realized the dimensional validities.

Procedure. The apparatus and procedure were the same as Experiments 1 and 2, except that no feedback was provided at the end of blocks that indicated the percentage correct. Participants were rotated through the conditions in the presumably quasi-random order in which they signed up for participation, which resulted in 48 participants in every condition except the high-salience relevant, no-irrelevant condition, which ended up with 49 participants.

Stimuli. The stimuli were words of high, moderate, or low salience, as determined by norms of concreteness, imagability, familiarity, and memorability. The selection of these stimuli was motivated by the need for strong control over stimulus salience that was demanded by this experiment. In previous attempts to generate the predicted interaction, we manipulated the height variation and segment variation of the geometric stimuli from Experiments 1 and 2. In two experiments, we found trends toward interactions as predicted, but the trends failed to reach statistical significance because the effect size was small. Power analyses indicated that hundreds of participants would need to contribute to each condition to make the power acceptably large. Instead of attempting to increase the effect size by further titrations of the relative variations of rectangle height and segment position, we changed to stimuli with better established saliences, namely, words.

In the long history of memory and learning experiments in which words were used as stimuli, it has been found that words of higher concreteness, imagability, and familiarity are usually easier to remember and associate (e.g., Christian, Bickley, Tarka, & Clayton, 1978). Because of such influences on word memorability, researchers have established normed values for hundreds of words on a variety of characteristic scales. We consulted the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981; available on-line at http://wapsy.psy.uwa.edu.au/uwa_mrc.htm) to obtain words of very high, average, or very low salience. As representative of high-salience words, we selected "boy" and "cat." These words have familiarity, concreteness, and imagability scores (respectively) of 1.19, 1.43, 1.56, and 0.95, 1.48, 1.55 *SD*s above the mean. As representative of moderate-salience words, we selected "peek" and "toll," which have values ±0.3 *SD*s of the mean on all three scales. For the low-salience words, we selected "nabob" and "witan," which have values on all three scales of more than 1.3 *SD*s below the mean. Thus, the alternative values of the high-salience dimension were "boy" and "cat"; the alternative values of the intermediate-salience dimension were "peek" and "toll"; and the alternative values of the low-salience dimension were "nabob" and "witan." In conditions when two words were presented on the same trial, they were shown one above the other, with their order randomly permuted.

As an additional, convergent manipulation of salience, the high-norm words were presented in all upper-case letters flanked by asterisks, the moderate-norm words were presented with initial capitals, and the low-norm words were presented in all lower-case letters. Instructions emphasized that, as an aid to learning, participants should try to imagine the referent of the word as they were learning. The complete text of the instructions is provided in Appendix A.

Associative learning experiments recently conducted (but as yet unpublished) in our laboratory showed that these words did indeed vary significantly, and consistently, in their ease of associating with deterministic categorical labels. Thus, we were encouraged that strong effects of salience would also appear in NMCPL. Moreover, these experiments would provide additional data regarding the generalizability of classic effects in NMCPL, originally explored using nonverbal stimuli.

The concept of stimulus salience does not yet have a conventional operationalization. Informally, a stimulus is salient to the extent that it is easy to encode or process, easy to distinguish from other stimuli in the same context, and easy to associate with outcomes. Words such as "boy" and "cat" satisfy these informal criteria, insofar as these words are highly practiced, have easily visualized referents, are semantically very rich and distinct from each other, and are easily associated with other words. Words such as "nabob" and "witan," on the other hand, might be distinctive and hence salient insofar as they are unusual, but because most people do not know what these words mean, the words are not as easy to encode, they do not evoke strong images, they are not semantically distinct from each other, and, for whatever reasons, they are not as easily associated with other words. Thus, our choice of stimuli specifically emphasized salience *qua* distinguishability and associability.

Results and Discussion

Results are shown in the top panel of Figure 15. Visual inspection clearly suggests that the RASHNL predictions were confirmed: When an irrelevant attribute of intermediate salience is added to the relevant attribute, the reduction in utilization of the relevant attribute is much greater for the low-salience relevant attribute than for the high-salience relevant attribute.

Statistical analysis verifies the reliability of these conclusions: Because the mean learning curves changed relatively little after Block 3 of training, we computed each participant's mean utilization of the relevant dimension, collapsing across Blocks 4 through 10. The distributions of utilizations were noticeably nonnormal, so all scores were converted to ranks (Conover & Iman, 1981). Even after this conversion, the lowest utilization group had notably smaller variance than the other groups, so the ranked data were analyzed using tests that did not pool variances. Despite these conservative steps, which typically reduce the power of the

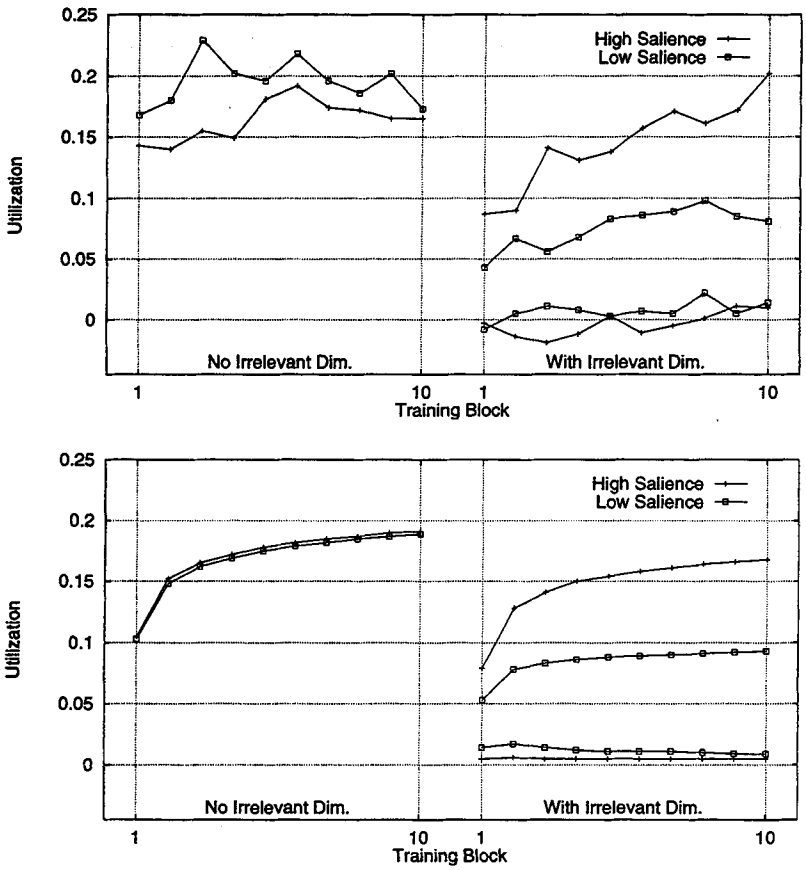

Figure 15. Upper graph: results of Experiment 3. Lower graph: best fit of RASHNL (Rapid Attention Shifts 'N' Learning), fit simultaneously to the results of Experiment 4. The abscissa is subdivided into two recurrences of Training Blocks 1–10, corresponding to the irrelevant dimension being absent or present. Within the right panels of each graph, the two curves marked with "+" symbols indicate the utilizations of the two cues when the relevant dimension (Dim.) had high salience; the curve near zero is for the irrelevant cue, and the higher curve is for the relevant cue. The two curves marked with square symbols indicate the utilizations of the two cues when the relevant dimension had low salience. Again, the curve near zero is for the irrelevant cue, and the higher curve is for the relevant cue.

tests, there was a significant interaction contrast, $t(189) = 2.08$, $SSE = 15.70$, $p = .039$, for adjusted $df = 179.7$, indicating that the effect of adding an irrelevant dimension was significantly different depending on the salience of the relevant dimension. This interaction was due largely to the small utilization of the low-salience relevant dimension in the presence of the intermediate-salience irrelevant dimension. Thus, the simple effect of salience, when the irrelevant dimension was present, was significant, $t(94) = 2.71$, $SSE = 10.19$, $p = .008$, for adjusted $df = 91.5$, and the simple effect of adding an irrelevant dimension, for the low-salience relevant dimension, was also significant, $t(94) = 3.18$, $SSE = 11.22$, $p = .002$, for adjusted $df = 85.7$. There was no significant simple effect of adding an irrelevant dimension to the high-salience relevant dimension, $t(95) = 0.27$, $SSE = 11.00$. When no irrelevant dimension was present, the unexpectedly higher utilization of the lower salience dimension can be discounted, as there was nowhere near a significant difference between mean utilizations of the high- and low-salience dimensions in the absence of an irrelevant dimension, $t(95) = 0.43$, $SSE = 11.95$. The trend toward a difference is mitigated further when considering the ranks themselves (as opposed to the raw utilizations plotted in Figure 15): The mean rank of the low-salience utilization was 109.2, $SD = 63.0$, and the mean rank of the high-salience utilization was nearly the same, at 104.1, $SD = 54.3$. All these conclusions are the same if "raw" utilizations are used instead of their rank-transformed values.

Other aspects of the data echo the results of Experiments 1 and 2. The learning curves are already at a fairly high level within Block 1 of training and rise relatively little after Block 3 of training. Within each condition, there is large between-subject variation in utilization of the relevant dimension. Scatterplots of individual utilizations are not included here, because they appear much like those shown in the leftmost panels of Figure 5.

Experiment 4: Salience of an Irrelevant Cue Affects Utilization

As discussed in a previous section, a second novel prediction of RASHNL is that the deleterious effect of adding an irrelevant dimension should be stronger when the irrelevant dimension is of high salience than when it is of low salience. This prediction directly conflicts with conclusions reached by Edgell (Edgell et al., 1992, p. 587; Edgell et al., 1996, p. 1477), who found trends but no significant effect of the salience of irrelevant cues. We now report results of an experiment that confirmed this prediction. In this experiment, an irrelevant cue of either high or low salience was added to a relevant cue that had moderate salience. The utilization of the moderate-salience cue was predicted to be higher when the irrelevant cue had low salience than when the irrelevant cue had high salience.

Method

Participants. A total of 137 students from introductory psychology courses at Indiana University volunteered for partial course credit.

Design. There was one relevant dimension of .2 validity, with moderate salience. It was accompanied by an additional, irrelevant dimension (.0 validity) of either high or low salience. Thus, the experiment comprised two conditions. As in Experiments 1, 2, and 3, training consisted of 10 blocks of 40 trials. Each block exactly realized the dimensional validities.

Procedure. The apparatus and procedure were the same as in Experiment 3. The instructions were identical to those of Experiment 3, and the full text is reported in Appendix A. Participants were rotated through the two conditions in the presumably quasi-random order in which they signed up for participation (i.e., every other participant was assigned to the low-salience condition), which resulted in 68 participants in the low-salience condition and 69 in the high-salience condition.

Stimuli. The stimuli were the same as in Experiment 3, that is, words of high, moderate, or low salience, as determined by published norms of concreteness, imagability, familiarity, and memorability.

Results and Discussion

Results are shown in the top panel of Figure 16. Visual inspection clearly suggests that the RASHNL predictions were confirmed: When the irrelevant attribute has high salience, the utilization of the relevant cue is less than when the irrelevant attribute has low salience.

Statistical analysis verifies the reliability of these conclusions: Because the mean learning curves changed relatively little after Block 3 of training, we computed each participant's mean utilization of the relevant dimension, collapsing across Blocks 4 through 10. The distributions of utilizations were noticeably skewed, so all scores were converted to ranks (Conover & Iman, 1981). The mean rank utilization for the high-salience irrelevant dimension was significantly less than the mean rank utilization for the low-salience irrelevant dimension, $t(135) = 2.32$, $SE = 6.68$, $p = .022$. This conclusion is the same if "raw" utilizations are used instead of their rank-transformed values, $t(135) = 2.77$, $SE = 0.022$, $p = .006$, for unequal-variance adjusted $df = 123.2$.

Other aspects of the data echo the results of our previous experiments. The learning curves are already at a fairly high level within Block 1 of training and rise relatively little after Block 3 of training. Within each condition, there is large between-subject variation in utilization of the relevant dimension. Scatterplots of individual utilizations are not included here, because they appear much like those shown in the leftmost panels of Figure 5.

Why did we observe an effect of the salience of an irrelevant dimension, when Edgell and colleagues did not? We believe it is simply a matter of statistical power. Edgell et al. did find some trends in their data consistent with our results, but the trends did not reach statistical significance. Recall that there is extremely large variance among participants in these experiments, so that statistical detection of differences in mean utilization demands both a large sample size and as large an effect size as possible. The stimuli we used (i.e., words of very different concreteness, memorability, and visual impact) were intended to be of extremely different salience, thereby generating, we hoped, a relatively large effect size. Despite these efforts, our data indicate an effect size of only 0.37. It is possible that the geometric stimuli used by Edgell et al. did not have salience differences as large.

Implication for Theory of Cue Confusion in STM

Edgell et al. (1992, 1996) argued that salience affects utilization by means of the memorability of cues in STM, and not via competition for attention. According to their memory hypothesis, a low-salience cue suffers more confusions in STM, and hence its effective validity is lower than its actual validity. Edgell et al. (1996) stated:

> Note that assuming memory errors are random and unbiased, they cannot change the perceived validity of an irrelevant cue dimension, because the association will still be 50/50. Hence, different physical representations of the irrelevant dimension would not have an effect on the utilization of the relevant dimension if the memory error hypothesis is correct, because the perceived validity of the irrelevant dimension would not change. However, if the alternative conspicuousness hypothesis is correct, then a more salient physical representation of the irrelevant dimension would draw more attention from the relevant dimension, causing it to be less utilized, than would a less salient physical representation of the irrelevant dimension. (p. 1475)

Our present findings therefore appear to disconfirm the memory hypothesis, at least in the form hypothesized by Edgell et al. (1996). RASHNL is not entirely inconsistent with the general idea of salience influencing effective validity in STM, however. What RASHNL adds to the memory hypothesis is the notion that salience competes for attention *before* cues are encoded in STM. Therefore, the salience of an irrelevant dimension does not necessarily alter its effective validity in STM, but the salience *does* alter the encoding of competing, relevant dimensions in STM.

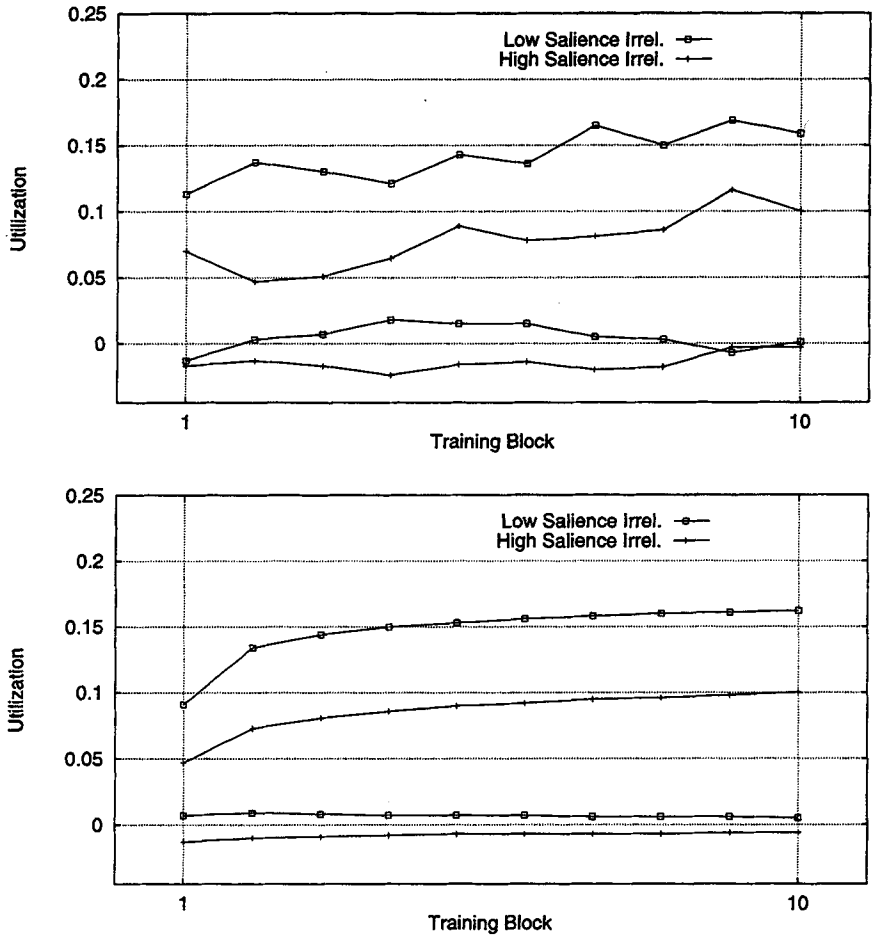

Figure 16. Upper graph: results of Experiment 4. Lower graph: best fit of RASHNL (Rapid Attention Shifts 'N' Learning), fit simultaneously to the results of Experiment 3. Within each graph, the two curves marked with "+" symbols indicate utilizations of the two cues when the irrelevant (Irrel.) cue had high salience; the curve near zero utilization is for the irrelevant cue, and the higher curve is for the relevant cue. The curves marked with square symbols indicate utilizations of the two cues when the irrelevant cue had low salience. Again, the curve near zero utilization is for the irrelevant cue, and the higher curve is for the relevant cue.

Fit of RASHNL to Results of Experiments 3 and 4

The results confirmed the qualitative predictions of RASHNL, when using the parameter values that best fit results from Experiments 1 and 2, which used geometric figures as stimuli. As a further test of the model, it was quantitatively fit simultaneously to the results of Experiments 3 and 4, which include the 100 data points plotted in the upper panels of Figures 15 and 16. The model was trained on the same sequences seen by the 330 participants. The best fitting predictions are shown in the lower panels of Figure 15 and 16, where it can be seen that the fit is excellent. The minimized RMSD is .018, with parameter values of $\phi = 2.91$, $\lambda_w = 0.208$, $\lambda_\alpha = 45.7$, $\rho = 0.483$, $P = 6.24$, $c = 8.86$, and saliences of 0.866, 1.00 (fixed), and 1.20 for the three different levels of salience. These parameter values are similar in magnitude to those that best fit the data from Experiments 1 and 2.

When there were two dimensions present, RASHNL tended to shift most of its attention to the more salient dimension, whether or not that dimension was valid. Thus, when the irrelevant dimension had higher salience than the valid dimension, the valid dimension was not utilized very strongly because it did not garner much attention. When the irrelevant dimension had lesser salience than the valid dimension, the valid dimension was utilized almost as well as when there was no irrelevant dimension, because the irrelevant dimension did not very strongly detract from the relevant dimension.

To reiterate our predictive logic, the parameter values from Experiments 1 and 2 were used to make qualitative predictions for the situations of Experiments 3 and 4 (and a number of other situations reported above). We assumed that the different stimuli used in these various experiments were dissimilar enough to demand different exact parameter

values for best quantitative fits by RASHNL but that the different experiments' stimuli were similar enough that the various best fitting parameter values should yield the same qualitative predictions. The qualitative predictions derived from the best fit to Experiments 1 and 2 were confirmed by our new results in Experiments 3 and 4. The model was then more stringently checked for its ability to quantitatively fit these new data, allowing different parameter values for these different stimuli. The fit was excellent. New parameter values were justified for this quantitative fit because of the stimulus differences, which could engender different saliences, different attentional capacity demands, different attentional separabilities, and so forth. The logic used here might be compared with the *logic* of predictions from, say, Newton's theory of gravity: The theory predicts that two objects on any planet should fall qualitatively in the same way (e.g., with the same acceleration). But quantitative predictions from the theory depend on planet-specific parameter estimates (e.g., a prediction of the magnitude of acceleration depends on an estimate of the planet's mass).

Utilization of Irrelevant Cues and Apparent Base-Rate Neglect

In all the cases we have considered so far, the cues were uncorrelated. That is, each combination of cues occurred equally often. The literature on probabilistic category learning has also recently highlighted cases of correlated cues, such that different combinations of cues occur with different frequencies. We now apply RASHNL in this situation.

Gluck and Bower (1988b) reported one case in which the cues were correlated and also the two categories had different base rates. The main point of their often-cited results was that for a particular cue, the true probability of the rare disease, given that cue alone, was 50%, but human participants chose the rare disease significantly more often than 50% of the time. This phenomenon has been dubbed "apparent base-rate neglect," because it might be the result of insufficient weighting of category base rates in classification. Gluck and Bower (1988b) explained the effect in a different way, as the consequence of error-driven learning of associative weights between cues and categories, formalized in their component-cue model.

Kruschke (1996a) argued that apparent base-rate neglect is actually caused by the additional influence of rapid, error-driven attention shifts between cues and not solely by error-driven associative weights. The model proposed by Gluck and Bower (1988b) had no such attention shifts. Experiment 4 of Kruschke (1996a) was a variation of Gluck and Bower's (1988b) design, in which apparent base-rate neglect was observed, but which could not be adequately fit by the component-cue model. Kruschke (1996a) showed that an extended version of the component-cue model, called ADIT, that incorporated rapid shifts of attention fit the data much better. Thus, the component-cue model was initially believed to account for base-rate neglect but was subsequently shown not to generalize to situations with different conditional probabilities. Rapid attention shifting was added to the model to account for the effects. RASHNL's predecessor, ALCOVE, was also initially believed to account for base-rate neglect (Kruschke, 1992; Nosofsky et al., 1992) but was subsequently shown not to generalize to different training sequences (Lewandowsky, 1995). Therefore it is appropriate to examine whether the addition of rapid attention shifting to ALCOVE, in the form of RASHNL, can account for the results of Experiment 4 of Kruschke (1996a).

Participants in the experiment saw, on each trial, a list of three symptoms, and the learner had to diagnose the case as one of two possible diseases. The three symptoms occurred in mutually exclusive pairs, thereby abstractly structured as three binary-valued dimensions. For example, symptom dimension A (sA) might have one value (sA = 1) of "hair loss" and an alternative value (sA = 2) of "stomach cramps." Symptom dimension B (sB) might have one value (sB = 1) of "blurred vision" and an alternative value (sB = 2) of "swollen tonsils." Thus, nothing in the symptom words themselves indicated dimensional membership. These stimuli were designed specifically to obscure the mutual exclusivity of the alternative values, so that learners would not use meta-associative strategies such as the reasoning, if Value 1 of sA indicates one disease, then Value 2 of sA must indicate the other disease. Thus, the six symptoms are best represented psychologically as six separate attributes, rather than as three binary dimensions. Across the 200 training trials, the common disease occurred 75% of the time, whereas the rare disease occurred 25% of the time.

The conditional probabilities of the symptoms, given the diseases, are shown in Table 3. Two aspects are important to point out. First, the conditional probabilities were designed so that when given sB = 2 alone, or sC = 2 alone, the probability of the rare disease is exactly 50%. If people choose the rare disease more than the common disease in these cases, then we have apparent base-rate neglect. Second, the first symptom, sA, is not correlated with the diseases, so, from a normative perspective, its value should have no influence on people's diagnoses. Table 4 shows the normative probabilities of the common disease for all possible symptom combinations.

After training, participants were shown test cases that included 27 symptom combinations, generated by crossing each of the symptoms' alternative values, or the absence of either value. In particular, one case was the null case, in which no symptoms were present. Other cases included only

Table 3
Conditional Probabilities of Symptoms Given Diseases for Experiment 4 Reported by Kruschke (1996a)

Symptom value	Disease	
	Common	Rare
sA = 2	.8667	.8667
sA = 1	.1333	.1333
sB = 2	.2000	.6000
sB = 1	.8000	.4000
sC = 2	.1333	.4000
sC = 1	.8667	.6000

Note. sA = 2 indicates Value 2 of symptom dimension A, sA = 1 indicates Value 1 of symptom dimension A, and so on.

Table 4
Proportion of Choices for the Common Disease for All 27 Cases From the Test Phase of the Experiment With Correlated Cues by Kruschke (1996a)

Symptom			Source of choice proportion			
sA	sB	sC	Norm	Human	RASHNL	ECCM
Null case						
0	0	0	.750	.768	.835	.747
One-symptom cases						
0	0	1	.812	.923	.921	.823
0	0	2	.500	.415	.377	.454
0	1	0	.857	.894	.918	.870
0	2	0	.500	.324	.338	.385
1	0	0	.750	.606	.644	.710
2	0	0	.750	.859	.818	.717
Two-symptom cases						
0	1	1	.897	.923	.962	.937
0	1	2	.667	.582	.652	.690
0	2	1	.591	.648	.618	.554
0	2	2	.250	.218	.235	.253
1	0	1	.812	.843	.773	.823
1	0	2	.500	.486	.453	.513
1	1	0	.857	.796	.753	.865
1	2	0	.500	.345	.449	.463
2	0	1	.812	.859	.932	.842
2	0	2	.500	.577	.509	.490
2	1	0	.857	.929	.930	.883
2	2	0	.500	.592	.446	.432
Three-symptom cases						
1	1	1	.897	.852	.824	.929
1	1	2	.667	.567	.589	.722
1	2	1	.591	.486	.579	.618
1	2	2	.250	.333	.416	.335
2	1	1	.897	.915	.975	.947
2	1	2	.667	.754	.723	.726
2	2	1	.591	.669	.687	.611
2	2	2	.250	.296	.355	.293

Note. sA heads the value of symptom dimension A, and so forth. A symptom value of 0 indicates that the symptom was absent. Norm = normative probability; Human = human choice proportions; RASHNL = predictions of Rapid Attention Shifts 'N' Learning; ECCM = enhanced component-cue model.

one symptom. For all these cases, participants were instructed to make their best guess based on what they had learned before.

Results from the test cases are shown in Table 4. For the null case, people guessed the common disease just slightly more often than the true base rate. For the single-symptom cases, people exhibited apparent base-rate neglect, choosing the common disease significantly less than 50% for sB = 2 and for sC = 2.

The single-symptom cases also clearly reveal that the first symptom was utilized by participants despite its normative irrelevance; thus, the preference for the common disease is much lower when sA = 1 than when sA = 2. This utilization of the irrelevant first symptom can also be observed in the training cases (i.e., the three-symptom cases at the bottom of Table 4). By considering the mean proportion of common choices when sA = 2, subtracted from the mean proportion of common choices when sA = 1 (i.e., the mean of the last four rows of Table 4 minus the mean of the immediately preceding four rows), we produce an indicator of utilization of the first symptom. For the normative probabilities, this difference is, of course, zero. For the human data, the difference is +0.099. This utilization of the first symptom was also strongly present throughout training, not just in testing.

RASHNL was fit to the data by being trained on the same 71 distinct sequences as the human participants saw. The six symptoms were represented by six input nodes, with values of 0 or 1 for absent or present symptoms, respectively. The best fitting predictions of RASHNL are shown in Table 4. RASHNL successfully shows preference for the common disease when no symptoms are present (i.e., for the null case), yet at the same time it also shows apparent base-rate neglect for both one-symptom cases (sC = 2 and sB = 2). RASHNL also successfully shows robust utilization of the irrelevant first symptom. This utilization can be seen for the single-symptom cases, where $p(C|sA = 1) = .644$ but $p(C|sA = 2) = .818$. The utilization of the irrelevant first symptom can also be seen for the three-symptom (training) cases: The mean choice proportion when sA = 2 subtracted from the mean choice proportion when sA = 1 is 0.083.

Further inspection of Table 4 reveals a few cases for which the fit by RASHNL is noticeably imperfect; the minimized RMSD was .0586, for parameter values of $\phi = 4.69$, $\lambda_w = 0.661$, $\lambda_\alpha = 52.0$, $\rho = 0.443$, $P = 29.9$, $c = 9.89$, and saliences fixed at 1.0. Qualitatively, however, the fit is good enough to lend some support to the model, especially when (a) contrasted with the notably poorer fit, RMSD = .0810, by an enhanced component-cue model (ECCM), described below, and (b) compared with the less-than-perfect best fit yet achieved by any model, RMSD = .0435, for ADIT (Kruschke, 1996a).

RASHNL successfully utilizes the irrelevant symptom because of rapidly shifting attention. During training, the three most frequently occurring cases, accounting for 80% of the training trials, have sA = 2, with the correct response being the common disease for 79% of these cases. Therefore, early in training, Value 2 of symptom sA becomes associated with the common disease. Moreover, of the rare cases, 87% also have sA = 2. Therefore, when a case of the rare disease occurs, it will typically contain Value 2 of symptom sA, and because this symptom is already associated with the common disease, *attention will shift away from this symptom in order to reduce error.* Thus, Value 2 of symptom sA remains associated with the common disease, resulting in significant utilization despite its normative irrelevance.

The performance of RASHNL can be contrasted with the best fit of ECCM, described by Kruschke (1996a). The ECCM has the original component cue model of Gluck and Bower (1988b) as a special case when certain parameters are fixed at zero, so if the enhanced model does not fit the data, neither does the original component cue model. The enhancements included the base-rate bias mechanism used by ADIT, the choice probability mapping function used by ADIT and RASHNL, and the feature-expectancy learning mechanism proposed by Shanks (1992). The best fitting predictions of this model are also shown in Table 4, which yield RMSD = .0810. The ECCM exhibits apparent base-rate neglect when

sC = 2 and when sB = 2, but it fails to show any utilization of the first symptom. This failure is evident by examining the one symptom cases, where the ECCM predicts that $p(C|sA = 1) = .710$ and $p(C|sA = 2) = .717$, a negligible difference. The failure to utilize the first symptom is also shown for the three-symptom training cases: The mean choice proportion for sA = 2 subtracted from the mean choice proportion for sA = 1 is 0.099 for humans but −0.007 for ECCM.

The fit of RASHNL to these data reveals the important fact that RASHNL robustly exhibits apparent base-rate neglect, despite the fact that its predecessor, ALCOVE, did not. The fit to these data also demonstrates that RASHNL utilizes a normatively irrelevant dimension, just as humans do, whereas ECCM did not.

General Discussion

Summary

We have shown that a wide variety of effects observed in multiple-cue probability learning can be accounted for by a model, named RASHNL, that implements three basic principles: rapid, error-driven shifting of limited-capacity attention; similarity-based exemplar representation; and annealing of learning and shift rates.

The various mechanisms of RASHNL work together and simultaneously and are not invoked separately to account for different effects. However, because each explanatory principle is parametrically formalized, its influence on the behavior of the model can be assessed. This correspondence of explanatory principles with formal mechanisms gives RASHNL clear explanatory power, and, despite its use of connectionist formalisms, the model does not suffer from the explanatory opacity ascribed to some connectionist models (McCloskey, 1991). In these other connectionist models, it is sometimes unclear how generic, low-level computational principles produce specific high-level behaviors.

Our approach to demonstrating the model's abilities was to fit RASHNL to data from two new experiments that partially replicated and modestly extended previous experiments in the literature and then to use the best fitting parameter values for making qualitative predictions of RASHNL for other situations previously reported in the literature. These predictions were qualitative (not quantitative), because the other situations used different stimuli, procedures, and apparatus, which could affect the specific dimensional saliences, attentional capacity, learning rates, and so forth. The other situations were similar enough, however, that we presumed that they could be qualitatively modeled with the parameter values that best fit Experiments 1 and 2. Thus, RASHNL not only explains the data that we fit directly but also predicts other effects. Moreover, we presented two new predictions—that the effect of salience should interact with the number of irrelevant cues, and the salience of an irrelevant cue should affect utilization—and we confirmed the predictions empirically in Experiments 3 and 4. As an additional test of the model, we examined its ability to quantitatively fit these new data, using freely estimated parameter values to accommodate the different stimuli; the resulting fit was very good. Finally, we also showed that RASHNL utilizes an irrelevant cue and exhibits apparent base-rate neglect, just as human learners did in Experiment 4 of Kruschke (1996a). Accounting for all these phenomena has been very difficult for other models.

In the remainder of the article we discuss possible mechanisms for adaptive learning rates, cases of cue competition observed for metric cues, and prospects for a theory of learning that spans multiple species. Finally, we recapitulate the rationality of rapid attentional shifts.

Annealing and Adaptive Learning Rates

Annealing of learning and of attentional shifts has been essential to our account. We think of annealing as an adaptive response to unreliable, inconsistent error signals. That is, if, on successive exposures to a stimulus, the correct response varies from trial to trial, then the error should be discounted, because it is unreliable. This discounting of error implies that whatever is learned early, before learning rates are much reduced, remains relatively "frozen" into place.

We used a *fixed* annealing schedule for NMCPL experiments with *stationary* probabilities. Fixed annealing schedules are not appropriate for nonstationary environments such as experiments with delayed introduction of information. The real world also presents frequently changing contexts and contingencies, which people continue to learn about. To address these nonstationary environments, there is needed some sort of adaptive learning rate or adaptive discounting of error.

Ashby, Alfonso-Reese, Turken, and Waldron (1988, Appendix 1), also used a fixed annealing schedule in their COVIS [competition between verbal and implicit systems] model of category learning, but these researchers had to modify the annealing depending on the structure of the task and the performance of the model. In particular, the annealing was applied only to probabilistic structures, not to deterministic structures, and the annealing schedule was reset if performance by the model was too poor. These manipulations of the annealing schedule again suggest the need for an adaptive learning rate or adaptive discounting of error.

Previous research on adaptive learning rates includes work by Sutton and his collaborators and successors (e.g., Fang & Sejnowski, 1990; Gluck, Glauthier, & Sutton, 1992; Jacobs, 1988; Nosofsky et al., 1994; Sutton, 1992). The fundamental idea of this approach is that learning rates should be adjusted to reduce error. This idea is expressed mathematically as gradient descent on error with respect to the learning rates, just as learning of associative weights is governed by gradient descent on error with respect to the associative weights.

The basic formula for adjusting learning rates by gradient descent on error is easy to derive. Suppose that at time (or trial) τ we are interested in adjusting the learning rate $\lambda(\tau)$ for a variable $w(\tau)$. This variable, w, could be an associative weight, or an attention strength, or any other variable in the

model that is adjusted by learning. Suppose moreover that the variable is learned by gradient descent on error, $E(\tau)$, which in turn is some function, f, of the variable $w(\tau)$: $E(\tau) = f(w(\tau))$. Learning of w by gradient descent on error E means that

$$w(\tau) = w(\tau-1) - \lambda(\tau)\frac{\partial E}{\partial w}(\tau-1), \tag{10}$$

where the learning rate λ is indexed by τ instead of by $\tau - 1$ to indicate that the learning rate λ is updated before the variable w is updated. Equation 10 indicates that $w(\tau)$ is a function of $\lambda(\tau)$ and implies that

$$\frac{\partial w(\tau)}{\partial \lambda(\tau)} = -\frac{\partial E}{\partial w}(\tau-1). \tag{11}$$

We then compute the change in the learning rate as follows:

$$\begin{aligned}\Delta\lambda(\tau) &= -\epsilon\frac{\partial E}{\partial \lambda}(\tau)\\ &= -\epsilon\frac{\partial f(w(\tau))}{\partial \lambda(\tau)}\\ &= -\epsilon\frac{\partial f(w(\tau))}{\partial w(\tau)}\frac{\partial w(\tau)}{\partial \lambda(\tau)}\\ &= \epsilon\frac{\partial E}{\partial w}(\tau)\frac{\partial E}{\partial w}(\tau-1),\end{aligned} \tag{12}$$

where ϵ is a constant of proportionality, that is, a meta-learning rate for the learning rate. Equation 12 indicates that the learning rate λ for w is increased if the changes in w are of the same sign from one trial to the next, but the learning rate λ is decreased if the changes in w are of opposite signs on successive trials. Equation 12 is called the "delta-delta" rule (Jacobs, 1988).

The delta-delta rule suffers some difficulties in practical applications, as described by Jacobs (1988). Variations and extensions have been developed by a number of researchers, including Fang and Sejnowski (1990), Jacobs (1988), and Sutton (1992). All these variations retain the basic theme that learning rates should increase if changes are consistent from trial to trial (as can occur in a deterministic situation), but learning rates should decrease if changes are inconsistent from trial to trial (as occurs in probabilistic situations). Different approaches to parameter adjustment have been described by Almeida et al. (1998), Darken and Moody (1992), and Sompolinsky et al. (1995), among many other investigators in this active field of research.

Instead of approaching learned nonlearning with adaptive learning rates, an alternative approach is adaptive discounting of error. We are not aware of any previous publications in which this approach is taken. The central idea of this novel approach is that the learner can differentially attend to different sources of error, just as the learner can differentially attend to different cues. In this scheme, the learner can shift attention away from unreliable sources of error, just as the learner can shift attention away from unreliable cues. At the time of this writing, we have not yet given this approach a thorough formal treatment, and we leave the idea as a plausible possibility.

There is yet a third possibility for how a learner might handle unreliable error signals. Rather than merely reducing the learning rate for error-driven changes, or reducing attention to error, the error signal could be transformed into a more reliable signal. For example, the learner might transform inconsistent trial-by-trial teacher signals into consistent, long-run-average teacher signals. This transformation to long-run averages is inefficient, because it takes multiple trials to compute and is only as reliable as the sample size (number of trials) over which the average is computed. Such a transformation is also computationally costly, because it needs to be conditionalized on stimulus cues and cue combinations. Attentional capacity constraints and shifts could limit which of these conditional means are computed; hence, this approach does not eliminate the role of attentional shifts. Determining all of the conditional mean teacher signals is tantamount to learning the mapping from cues to outcomes, and so it is really no solution to the learning problem, unless there are actually two learning problems: one for determining the conditional mean teacher signals, and one for inferring what underlying mapping generated these signals. If a learner took this approach to handling unreliable teacher signals, then he or she could show "maximizing" instead of probability matching, because the learner could infer a *deterministic* underlying mapping from cues to outcomes that generates a stable long-run-average teacher. These speculations await exploration in future research.

Metric Cue Probability Learning

The NMCPL framework uses discrete cues and discrete outcomes, but cue competition effects are also observed in situations with continuously valued metric cues and outcomes. RASHNL is equipped to deal with metric cues but is not currently formulated to address metric outcomes. Even for discrete outcomes, RASHNL cannot fully address extrapolation beyond the domain of the trained values (Delosh, Busemeyer, & McDaniel, 1997; Erickson & Kruschke, 1998). Future models that address situations with metric cues and outcomes would presumably also incorporate the three essential principles of RASHNL. In this section we briefly review two such situations of cue competition observed for metric cues, one case with discrete outcomes, the other case with metric outcomes.

Metric cues and discrete outcomes. In an experimental paradigm called the *randomization technique,* Ashby et al. (1998) used metric cues that predicted discrete, categorical outcomes. In this paradigm, the cue values for an instance of a category are drawn from a multivariate normal probability distribution; that is, the probability of a cue value, given a category, is normally distributed. Ashby et al. reported an

experiment in which the stimuli consisted of two metric cues (physically realized as line segments that varied in length and orientation). Both cues were normally distributed within each of two categories, with equal variances and zero covariances. On the first dimension, the two category means were separated by 0.74 *SD*s, and on the second dimension the category means were separated by 1.18 *SD*s.

For our purposes it is important to notice that the cue with the greater separation between category means is the more valid cue. To understand this, consider a case in which the two categories have the same mean on the first dimension. Because the marginal probability distributions on this dimension are identical for the two categories, this cue has zero validity. At the opposite extreme, consider a case in which two categories have an extremely large separation between their means on this dimension. Because in this case there is virtually no overlap between categories, this cue has very high validity. Thus, in the category structure used by Ashby et al. (1998), the second dimension, with category means separated by 1.18 *SD*s, was more valid than the first dimension, which had category means separated by 0.74 *SD*s.

Ashby et al. (1998) found that after 2,000 trials of training, people did not utilize the dimensions proportionally to their validities. Instead, people underutilized the less valid dimension relative to the more valid dimension. (Ashby et al. did not measure validity and utilization as described here, in the context of NMCPL. Instead, they reported the slope of the best fitting linear discriminant between the two categories. This slope is directly related to the relative validities or utilizations of the two categories.)

Ashby et al. (1998) interpreted this result in the context of their multiple-system theory of categorization, named COVIS, to stand for competition between verbal and implicit systems. Verbal rules in this model are formalized as thresholds on single dimensions; for example, the stimulus is in Category K if its value on Dimension A exceeds threshold Value V. In COVIS, even if the implicit system accurately learns the relative validities of the dimensions, the verbal system will tend toward a rule on the more valid dimension. When the results of the two systems are mixed, there will be a bias toward the more valid dimension, away from the less valid dimension, thereby accounting for the result.

Alternatively, the result can be construed simply as a case of overshadowing of a less valid cue by a more valid cue. In the competition for attention to dimensions, the more valid dimension tends to attract more attention, at the expense of reduced attention to the less valid dimension. Rapid shifts of attention, as posited by RASHNL, might also account for the dynamics of learning observed by Ashby et al. (1998). They reported that early in training, participants would utilize just one dimension, then just the other dimension, switching back and forth, but converging to stable utilizations later in training.

It would be straightforward to apply RASHNL to this case, but it would also be well beyond the intended scope of this article, which focuses on the NMCPL paradigm. We leave it to future research to explore simulations of RASHNL applied to the numerous experiments in the literature that have used the randomization technique.

Metric cues and metric outcomes. A clear example of cue competition among metric cues with metric outcomes was reported by Busemeyer et al. (1993a). Participants in their experiment learned to predict the height of a plant that was given various amounts of two growth hormones. The correct height was generated by a linear combination of the two hormone quantities plus normally distributed noise. A critical design feature was that the hormone quantities were uncorrelated across trials, so that the marginal validity of each cue was unaffected by the validity of the other cue. Busemeyer et al. (1993a) held constant the validity of one cue but set the validity of the second cue to a lower level in one condition and to a higher level in another condition. Utilization of the fixed-validity cue was significantly decreased when the other cue had higher validity.

In summary, not only is cue competition observed in situations with discrete cues and outcomes, but it has also been documented in situations with continuously varying metric cues and outcomes. Presumably, the same psychological principles are at work in all these situations, and future formalizations of the principles will address continuous metric cues and outcomes.

Comparison Across Species

We began this article by noting that cue competition is a ubiquitous phenomenon, exhibited not only by humans in a variety of situations but also by nonhuman animals, even honeybees. This commonality of behavior suggests that similar principles of learning may be involved across species, even if the specific biological mechanisms that implement the principles are quite different. In addition to the commonalities across species, obviously there are also many differences in the cognitive abilities of different species. In the interest of constructing a unified theory of learning, it is tempting to contemplate whether a single model (which formalizes a specific set of principles at the algorithmic, not implementational, level) can address behavior of various species by merely changing parameter values for different species. In particular, a variety of species might all learn with attention shifts, but different species might have different rates of attention shift and different rates of learning the shifts.

Previous researchers have contemplated this unified approach. Trabasso and Bower (1968) cautioned that

> a more critical question is whether one should even attempt to construct a theory that is proposed to be valid for animals and men except for variations in the parameters. Though an attractive strategy, it may simply be unrealistic to expect even moderate success along these lines. (p. 223)

Undeterred, Mackintosh (1969) compared probability learning among rats, birds, and fish and concluded that

> the simplest explanation, therefore, of the behavioural differences between rat, bird and fish, is to suggest that the three classes of animal differ in the extent to which they can learn to attend to a given cue when it is not consistently correlated with reinforcement. (pp. 148–149)

Recent work by Kruschke (1997a) shows that the model of attentional learning in animals proposed by Mackintosh (1975) is (very nearly) a special case of the model attentional learning in humans proposed by Kruschke (1996a). The fact that models motivated by different species turn out to be closely related adds encouragement to those who seek a unified theory of learning.

Rapid Attention Shifts Are Rational

This article also began with the assertion that it would seem rational or optimal to learn about partial correlations between cues and outcomes as they actually exist in the world, but instead it is the case that humans and other animals exhibit nonnormative utilization of cues. Have all these species thrived despite having irrational, suboptimal learning? Have the selective pressures of all these species' evolutionary niches been so benign that inaccurate learning goes unpunished in reproductive success? We believe the answer to these questions is "no." Instead, the learning behavior of these species is an evolutionarily adaptive solution to a constraint on learning other than long-run accuracy: the need for speed. By this we mean the need to learn a new association in as few exposures as possible, without destroying previously acquired associations. An organism that learns quickly would probably possess a reproductive advantage over competitors that learn less quickly.

Rapid, error-driven shifting of attention is a method to achieve the goal of speedy learning. This is because error-driven shifts of attention are tantamount to *reduction of interference* between different associations, and reduction of interference allows faster acquisition of new associations without destroying old associations (Kruschke, 1997a, 1997b). Shifts of attention reduce interference, because attention is shifted away from cues that cause error toward cues that reduce error. Which cues cause error? Those cues that are already associated with a response different from the one presently demanded. By shifting attention away from the cues previously associated with different outcomes, those associations are preserved and protected from overwriting. Which cues reduce error? Those that are already associated with the response presently demanded. By shifting attention to these cues already associated with the desired response, learning is speeded and redundant cues are reserved for future use.

The need for speedy learning is probably a constraint imposed on numerous species, and it may be the case that numerous species have evolved forms of rapidly shifting selective attention to address this need. Whereas the specific biological implementations of attention shifting may differ across species, the behaviors might have a functional commonality that can be captured in a single algorithmic formalization. Moreover, the behaviors that result from rapid attention shifts, such as overshadowing, are rational and nearly optimal with respect to the constraint of efficiency.

Conclusion

We conclude with a recapitulation of the main points. At the empirical level, Experiments 1 and 2 replicated previous results in the literature but also provided new details not previously reported, including extensive individual differences in utilization, and an interaction of salience and validity. New predictions of RASHNL, regarding competitive effects of the salience of an irrelevant cue, were verified by Experiments 3 and 4.

At the theoretical level, we have argued that the critical psychological mechanisms, needed to explain the panoply of effects addressed here, include rapidly shifting attention and annealed (or better yet, adaptive) learning rates. No previous model has been able to account for the range of effects addressed here, and no previous model has implemented these explanatory principles.

At the meta-theoretical level, the article suggests that the apparently irrational, nonnormative behaviors, shown by humans and other animals, are in fact natural consequences of highly adaptive solutions to the problem of fast learning. Many species presumably share the need to learn in as few trials as possible, without overwriting previously learned knowledge. Rapidly shifting attention is a functional solution to this need for rapid interference reduction.

References

Almeida, L. B., Langlois, T., Amaral, J. D., & Plakhov, A. (1998). *Parameter adaptation in stochastic optimization* [Online]. Available: FTP://146.193.2.131/pub/lba/papers/adsteps.ps

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

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.

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 414–432.

Björkman, M. (1967). Stimulus-event learning and event learning as concurrent processes. *Organizational Behavior and Human Performance, 2,* 219–236.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3–11.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993a). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science, 4,* 190–195.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993b). Cue competition effects: Theoretical implications for adaptive network learning models. *Psychological Science, 4,* 196–202.

Castellan, N. J. (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Performance, 9,* 16–29.

Castellan, N. J., & Edgell, S. E. (1973). An hypothesis generation model for judgment in nonmetric multiple-cue probability learning. *Journal of Mathematical Psychology, 10,* 204–222.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545–567.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition, 21,* 413–423.

Christian, J., Bickley, W., Tarka, M., & Clayton, K. (1978). Measures of free recall of 900 English nouns: Correlations with imagery, concreteness, meaningfulness, and frequency. *Memory & Cognition, 6,* 379–390.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505.

Conover, W., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35,* 124–129.

Darken, C., & Moody, J. (1991). Note on learning rate schedules for stochastic optimization. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems* (Vol. 3, pp. 832–838). San Mateo, CA: Kaufmann.

Darken, C., & Moody, J. E. (1992). Toward faster stochastic gradient search. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems* (Vol. 4, pp. 1009–1016). San Mateo, CA: Kaufmann.

Delosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986.

Edgell, S. E. (1978). Configural information processing in two-cue probability learning. *Organizational Behavior and Human Performance, 22,* 404–416.

Edgell, S. E. (1980). Higher order configural information processing in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance, 25,* 1–14.

Edgell, S. E. (1983). Delayed exposure to configural information in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance, 32,* 55–65.

Edgell, S. E. (1993). Using configural and dimensional information. In N. J. Castellan (Ed.), *Individual and group decision making* (pp. 43–64). Hillsdale, NJ: Erlbaum.

Edgell, S. E., Bright, R. D., Ng, P. C., Noonan, T. K., & Ford, L. A. (1992). The effects of representation on the processing of probabilistic information. In B. Burns (Ed.), *Percepts, concepts and categories* (pp. 569–601). New York: Elsevier Science.

Edgell, S. E., & Castellan, N. J. (1973). Configural effect in multiple-cue probability learning. *Journal of Experimental Psychology, 100,* 310–314.

Edgell, S. E., Castellan, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1463–1481.

Edgell, S. E., & Hennessey, J. E. (1980). Irrelevant information and utilization of event base rates in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance, 26,* 1–6.

Edgell, S. E., & Morrissey, J. M. (1987). Delayed exposure to additional relevant information in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 40,* 22–38.

Edgell, S. E., & Roe, R. M. (1995). Dimensional information facilitates the utilization of configural information: A test of the Castellan–Edgell and the Gluck–Bower models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1495–1508.

Edgell, S. E., Roe, R. M., & Zurada, J. M. (1993, November). *Connectionist modeling of learning in a probabilistic (decision making) environment.* Paper presented at the 34th Annual Meeting of the Psychonomic Society, Washington, DC.

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

Estes, W. K. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 89–128). New York: Academic Press.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37–64.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–576.

Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology, 47,* 225–234.

Fang, Y., & Sejnowski, T. J. (1990). Faster learning for dynamic recurrent backpropagation. *Neural Computation, 2,* 270–273.

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

Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

Gluck, M. A., & Bower, G. H. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Gluck, M. A., Glauthier, P. T., & Sutton, R. S. (1992). Adaptation of cue-specific learning rates in network models of human category learning. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 540–545). Hillsdale, NJ: Erlbaum.

Goldstein, I. L. (1973). Irrelevant information as a variable in complex displays. *Behaviorometric, 3,* 67–73.

Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis of variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin, 69,* 338–349.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1,* 295–307.

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

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

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science, 5,* 3–36.

Kruschke, J. K. (1993b). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: The psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego, CA: Academic Press.

Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 3–26.

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science, 8,* 201–223.

Kruschke, J. K. (1997a, November). *Attention in learning: Relating Mackintosh's (1975) theory to connectionist models and human categorization.* Paper presented at the Eighth Australasian Mathematical Psychology Conference, Perth, Australia.

Kruschke, J. K. (1997b). Selective attention in associative learning

[Review of the book *The Psychology of Associative Learning*]. *Journal of Mathematical Psychology, 41,* 207–211.

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for human associative learning* [Indiana University Cognitive Science Research Report No. 141]. Available: http://www.indiana.edu/~kruschke/kb95abstract.html

Kruschke, J. K., & Erickson, M. A. (1995, November). *Six principles for models of category learning.* Paper presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles, CA. Available: http://www.indiana.edu/~kruschke/psychonomics95-abstract.html

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review, 102,* 185–191.

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

Macho, S. (1997). Effect of relevance shifts in category acquisition: A test of neural networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 30–53.

Mackintosh, N. J. (1969). Comparative studies of reversal and probability learning: Rats, birds and fish. In R. M. Gilbert & N. S. Sutherland (Eds.), *Animal discrimination learning* (pp. 137–162). New York: Academic Press.

Mackintosh, N. J. (1970). Attention and probability learning. In D. I. Mostofsky (Ed.), *Attention: Contemporary theory and analysis* (pp. 173–191). New York: Appleton-Century-Crofts.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276–298.

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning and Behavior, 4,* 186–192.

March, J., Chamizo, V. D., & Mackintosh, N. J. (1992). Reciprocal overshadowing between intra-maze and extra-maze cues. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 45B,* 49–63.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387–395.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Mellers, B. A., & Biagini, K. (1994). Similarity and choice. *Psychological Review, 101,* 505–518.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248–277.

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2,* 416–421.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego, CA: Academic Press.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3,* 222–226.

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.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (Vol. 2, pp. 64–99). New York: Appleton-Century-Crofts.

Ruebeling, H. (1993). Pavlovian conditioning in human skilled motor behavior. *Integrative Physiological and Behavioral Science, 1,* 29–45.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209–237.

Shanks, D. R. (1991a). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R. (1991b). A connectionist account of base-rate biases in categorization. *Connection Science, 3,* 143–162.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science, 4,* 3–18.

Shapiro, M. S., & Bitterman, M. E. (1998). Intramodal competition for attention in honeybees. *Psychonomic Bulletin & Review, 5,* 334–338.

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

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

Sompolinsky, H., Barkai, N., & Seung, H. S. (1995). On-line learning of dichotomies: Algorithms and learning curves. In J.-H. Oh, C. Kwon, & S. Cho (Eds.), *Neural networks: The statistical mechanics perspective* (pp. 105–130). Singapore: World Scientific.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning.* New York: Academic Press.

Sutton, R. S. (1992). Adapting bias by gradient descent: An incremental version of delta-bar-delta. *Proceedings of the 10th National Conference on Artifical Intelligence* (pp. 171–176). Menlo Park, CA: MIT/AAAI Press.

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

Wagner, A. R. (1969). Stimulus validity and stimulus selection in associative learning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 90–122). Halifax, Nova Scotia, Canada: Dalhousie University Press.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76,* 171–180.

Wasserman, E. A. (1974). Stimulus-reinforcer predictiveness and selective discrimination learning in pigeons. *Journal of Experimental Psychology, 103,* 284–297.

White, H. (1989). Some asymptotic results for learning in single hidden-unit feedforward network models. *Journal of the American Statistical Association, 84,* 1003–1013.

Appendix A

Full Text of Instructions to Participants

The following instructions were presented on the participants' computer screens at the beginning of the experiments.

Instructions for Experiment 1

This experiment investigates how people learn to classify and categorize objects. You will be shown a series of rectangles with a line segment in each. The height of the rectangle and the position of the line segment may change from one presentation to the next.

Each combination of rectangle and line segment is labeled with a letter. Your job is to learn to predict the correct label for each combination of rectangle with line segment. After a rectangle and line segment appear, you should type in what you think is the correct label.

The rectangle and line are only probabilistically related to the label. Thus a given picture might only usually be an 'F' instead of always being an 'F'. So even though you can't always be correct, it is still possible to get up to 70 or 80 percent correct if you try hard. You should try to respond as accurately as possible.

The correct response will be shown after you make your response so you can evaluate your performance.

[Press the space bar to continue]

Periodically during the experiment, you will be given a short break and shown how well you have done over the last block of trials.

Please place all personal belongings out of the way, under the desk. Please do not listen to music or take notes during the experiment since it could affect the outcome.

You should try to respond as accurately as possible. Please do not leave until the computer says that the experiment is over.

[Press the space bar to continue]

Throughout the experiment, when you see a rectangle and line displayed, respond by pressing the key that corresponds to what you think is the correct label. At first, you'll just be guessing what letter to choose. There are, however, only two choices: the letter F and the letter J.

Remember because this is a probabilistic task, you should be able to get up to 70 or 80 percent correct.

Make sure you know where the F and J keys are.

[Press the space bar to continue]

If you have any questions, ask the experimenter now.

WAIT for the experimenter to close the curtain and leave the room.

[Press SPACE BAR to BEGIN the Experiment]

Instructions for Experiment 2

This experiment investigates how people learn to classify and categorize things. You will be shown a series of rectangles that vary in height. In addition you might be shown a vertical line segment inside of the rectangle that varies in position.

Your job is to learn to predict the correct label for each rectangle or combination of rectangle with line segment. When the stimulus is displayed, you should type in what you think is the correct label. The possible labels are 'F' and 'J'.

A given stimulus is only probabilistically related to a label. Thus, a given stimulus might only usually be an 'F' instead of always being an 'F'. So even though you can't always be correct, it is still possible to get up to 70 percent correct if you try hard. You should try to respond as accurately as possible.

The correct response will be shown after you make your response so you can evaluate your performance.

[Press the space bar to continue]

Periodically during the experiment, you will be given a short break and shown how well you have done over the last block of trials.

Please place all personal belongings out of the way, under the desk. Please do not listen to music or take notes during the experiment since it could affect the outcome.

You should try to respond as accurately as possible. Please do not leave until the computer says that the experiment is over.

[Press the space bar to continue]

Throughout the experiment, respond by pressing the key that corresponds to what you think is the correct answer. At first, you'll just be guessing what letter to choose.

After a while, you'll learn which stimuli tend to go with which labels. Remember that because this is a probabilistic task, you should be able to get up to 70 percent correct.

Make sure you know where the F and J keys are.

[Press the space bar to continue]

If you have any questions, ask the experimenter now. WAIT for the experimenter to close the curtain and leave the room.

[Press SPACE BAR to BEGIN the Experiment]

Instructions for Experiments 3 and 4

This experiment examines how people learn to make accurate medical diagnoses. You will be presented with many patients' case histories. For each case history you will be shown the symptoms the patient has, and you will be asked to choose which illness you think the patient has. After you make your diagnosis, you will be told the correct diagnosis. All you have to do is try to learn which symptoms tend to go with which illnesses so that you can make as many correct diagnoses as possible.

Re-read the previous paragraph if it is unclear. Then, press the space bar to continue.

There are two possible diseases that the patients have, and each patient has one and only one of the diseases. In order to keep things as straight-forward as possible, we'll simply label the diseases with letters F and J. For each case history, you indicate your diagnosis by pressing one of these two letters on the keyboard. You'll have up to 30 seconds to make your diagnosis for each case history. At first you will just be guessing, but after many cases your accuracy will improve.

Just as with real symptoms and real diseases, any particular symptom might not be a perfect predictor of the diseases. For example, a fever might indicate a flu, but might instead indicate a bacterial infection. The symptoms in this experiment will merely tend to indicate particular diseases. You cannot always be correct, but you can learn the tendencies and get up to 70% correct.

If any of the instructions on this screen are unclear, please re-read them now. Otherwise, press the space bar to continue.

Instead of real symptoms, such as 'sore throat' or 'head ache,' you will be shown simple words such as 'code' or 'fable.' Words that represent especially severe or critical symptoms will be presented in all capitals with asterisks, like this:

IMPORTANT

Words that represent moderate or standard symptoms will be presented in initial capitals, like this:

Moderate

Words that represent unimportant, inconsequential symptoms will be presented in all lowercase letters, like this:

inconsequential

[Press the space bar to continue.]

It will be faster and easier to learn the diseases if you try to create a visual image of the symptom words. Thus, for each symptom word, try to imagine what the word refers to, as you are learning the correct disease. For example, if you see the symptom word 'fable' with disease A, try to imagine or visualize a fable as you associate it with A.

To reiterate, your task is to learn which symptoms tend to go with which diseases, so that you can make as many correct diagnoses as possible. Important symptom words are shown like this: *IMPORTANT*, moderate symptom words are shown like this: Moderate, and unimportant symptom words are shown like this: inconsequential. Because the symptoms are not perfect indicators of the diseases, you cannot make correct responses for every case, but if you keep trying to learn which diseases tend to be indicated by which symptoms, you can get up to 70% correct.

[Press the space bar to continue.]

If you have any questions, please ask now. WAIT for the experimenter to close the cubicle curtain and leave the room. Press the space bar to begin the experiment.

Appendix B

Derivation of Formula for Attention Shift

In this Appendix, we derive the formula for rapid shifting of attentional gains, presented as Equation 7 in the main text. The formula expresses the change in the gain on Dimension A as gradient descent on error with respect to that attentional gain; that is, $\Delta\gamma_A = -\lambda_\gamma \partial E/\partial\gamma_A$, where λ_γ is a constant of proportionality.

We denote the vector of category node activations by $\mathbf{a}^{cat} = [\ldots, a_k^{\text{cat}}, \ldots]^T$. Similarly, the vectors of exemplar node activations and of attention node activations are denoted by $\mathbf{a}^{ex} = [\ldots, a_j^{\text{ex}}, \ldots]^T$ and $\alpha^{att} = [\ldots, \alpha_i, \ldots]^T$, respectively.

By the chain rule of vector calculus,

$$\frac{\partial E}{\partial \gamma_A} = \frac{\partial E}{\partial \mathbf{a}^{cat}} \frac{\partial \mathbf{a}^{cat}}{\partial \mathbf{a}^{ex}} \frac{\partial \mathbf{a}^{ex}}{\partial \alpha^{att}} \frac{\partial \alpha^{att}}{\partial \gamma_A}$$

$$= \begin{bmatrix} \cdots & -(t_k - a_k^{\text{cat}}) & \cdots \end{bmatrix} \begin{bmatrix} \vdots & \\ \cdots & w_{kj} \quad \cdots \\ \vdots & \end{bmatrix}$$

$$\times \begin{bmatrix} \vdots & \\ \cdots & a_j^{\text{ex}}(-c)\sigma_i|\psi_{ji} - a_i^{\text{in}}| \quad \cdots \\ \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \partial\alpha_i/\partial\gamma_A \\ \vdots \end{bmatrix},$$

where

$$\frac{\partial \alpha_i}{\partial \gamma_A} = \left(\begin{array}{l} \left(\sum_j \exp(\gamma_j)^P\right)^{1/P} \exp(\gamma_A)\kappa_{iA} \\ \quad - \left\{ \exp(\gamma_i) \frac{1}{P} \left(\sum_j \exp(\gamma_j)^P\right)^{(1/P)-1} \right. \\ \quad \left. \times P \exp(\gamma_A)^{P-1} \exp(\gamma_A) \right\} \end{array} \right) \Big/ \left(\sum_j \exp(\gamma_j)^P \right)^{2/P}$$

$$= \kappa_{iA}\alpha_A - \alpha_i \alpha_A^P \theta.$$

Combining these expressions yields Equation 7 in the main text.

Received July 8, 1998
Revision received February 24, 1999
Accepted March 2, 1999 ■