Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 6, 999–1019

0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.6.999

Limitations of Exemplar Models of Multi-Attribute Probabilistic Inference

Robert M. Nosofsky and F. Bryan Bergert
Indiana University

Observers were presented with pairs of objects varying along binary-valued attributes and learned to predict which member of each pair had a greater value on a continuously varying criterion variable. The predictions from exemplar models of categorization were contrasted with classic alternative models, including generalized versions of a “take-the-best” model and a weighted-additive model, by testing structures in which interactions between attributes predicted the magnitude of the criterion variable. Under typical training conditions, observers showed little sensitivity to the attribute interactions, thereby challenging the predictions from the exemplar models. In a condition involving highly extended training, observers eventually learned the relations between the attribute interactions and the criterion variable. However, an analysis of the observers’ response times for making their paired-comparison decisions also challenged the exemplar model predictions. Instead, it appeared that most observers recoded the interacting attributes into emergent configural cues. They then applied a set of hierarchically organized rules based on the priority of the cues to make their decisions.

Keywords: multi-attribute inference, configural cues, exemplar models, heuristics, response times

In tasks of pairwise probabilistic inference, observers are presented with pairs of alternatives that vary along multiple attributes. The observers’ task is then to judge, on the basis of the attribute information, which of the two alternatives is greater on some criterion variable. For example, in the well-known city-population task (Gigerenzer & Goldstein, 1996), an observer is presented with two cities and is asked to judge which has a greater population. The observer is presumed to have knowledge about various attributes of the cities, such as whether they have subway systems, whether they have professional basketball teams, and whether they are state capitals. Assuming that the observer recognizes the names of both cities, he or she is presumed to use the attribute information as a basis for judging the cities’ sizes. In the present article, we limit consideration to situations in which the stimuli are composed of binary-valued attributes but vary continuously along the criterion variable.

Although knowledge about the relation between the attributes and the criterion may be obtained in numerous ways, in this research we focus on cases in which it is obtained via induction over presented pairs of examples. Thus, on each trial, the observer is presented with a pair of multi-attribute alternatives and judges which of the two alternatives is greater on the criterion. Trial-by-trial feedback is provided. The feedback is discrete in the sense that the only information that is provided is the correct forced-choice alternative. Feedback is not provided regarding the precise quantitative value of the continuous criterion variable. Similar forms of learning operate in the real world in various situations. For example, a observer may learn about the strength of different sports teams by viewing the outcomes of their head-to-head competitions (Heit, Price, & Bower, 1994). Likewise, in presidential campaigns, voters gain knowledge of the strength of alternative candidates from the results of the primary elections (Estes, 1976).

Two of the major models of probabilistic inference include the take-the-best (TTB) model and the weighted-additive (WADD) model (Gigerenzer & Selten, 2001; Gigerenzer & Todd, 1999; Juslin, Jones, Olsson, & Winman, 2003; Lee & Cummins, 2004; Rieskamp & Otto, 2006). According to both models, observers obtain information during training concerning the diagnosticity of the individual attributes (for details, see The Formal Models section). According to TTB, observers order the cues according to their diagnosticity and compare the alternatives on an attribute-by-attribute basis until a discriminating cue is found. The observer then chooses the alternative with the positive value of that cue, that is, the one that points to the higher value of the criterion variable. According to the WADD model, the observer considers all of the cues that discriminate between the alternatives. The evidence in favor of Alternative A is found by summing the diagnosticities of the cues that favor A and likewise for the evidence in favor of Alternative B. The observer then chooses the alternative with the greater summed evidence.

A great deal of research has been devoted to contrasting TTB and WADD strategies and determining the conditions under which they operate (e.g., Bröder & Schiffer, 2003; Newell & Shanks, 2003; Payne, Bettman, & Johnson, 1988; Rieskamp & Otto, 2006). By contrast, there has been less work considering other possible modes of decision making in the probabilistic inference task. The central purpose of the present research was to investigate whether one of the major models of categorization processes may be applicable. By categorization, we mean the cognitive process in which observers learn to assign distinct objects into classes or groups. Categorization is related to the probabilistic inference task

Robert M. Nosofsky and F. Bryan Bergert, Department of Psychological and Brain Sciences, Indiana University.

This work was supported by National Institute of Mental Health Grant R01 MH48494. We would like to thank Matt Jones, Michael Lee, and Ben Newell for their helpful criticisms of earlier versions of this article.

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

in various ways. For example, in both tasks, information from multiple attributes carrying probabilistic evidence is integrated and forms the basis for subjects' decisions. A fundamental difference is that the probabilistic inference task requires observers to make more fine-grained decisions and to discriminate each of the individual alternatives along a continuous-valued criterion variable. By contrast, in typical tasks of categorization, multiple instances are assigned the same nominal-valued categorization response. Nevertheless, it is of interest to consider whether similar cognitive processes may mediate these forms of judgment (cf. Nosofsky, 1986).

In the present research, our central aim is to test an exemplar-based model of categorization in the context of the paired-comparison, probabilistic inference task. According to exemplar models (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986), people represent categories by storing individual exemplars in memory and classify objects on the basis of their similarity to these stored exemplars. A natural idea is that exemplar-storage processes may be involved in multi-attribute probabilistic inference. For example, if a test item is similar to numerous previously experienced exemplars that have a high value on the criterion variable, then an observer is likely to judge that the test item is high on the criterion as well. Indeed, there is evidence that similar forms of exemplar-based "reminding" operate in diverse forms of learning and instruction (e.g., Ross, 1989).

The idea that stored exemplars may drive probabilistic inference is not new (Dougherty, Gettys, & Ogden, 1999; Juslin & Persson, 2002). For example, according to the influential probabilities from exemplars (PROBEX) model of Juslin and Persson (2002), people estimate the value of a continuous quantity by a sequential retrieval of exemplars from long-term memory. These researchers demonstrated that the PROBEX model makes accurate normative predictions of continuous-valued attributes in real-world environments. In addition, the researchers demonstrated that PROBEX is psychologically descriptive by showing that it provides excellent quantitative accounts of a variety of continuous-valued point estimates, binary decisions, and subjective probability judgments made by humans. However, Juslin, Jones, et al. (2003) also demonstrated cases in which exemplar models failed to predict human performance in categorization tasks involving the presentation of continuous-valued feedback. In these tasks, the experimenters defined an additive integration rule for assigning a continuous criterion variable to each combination of binary-valued stimulus cues. In training conditions that provided precise continuous-valued feedback, it appeared that subjects learned the additive integration rule rather than relying on exemplar memory. Likewise, researchers in the domain of continuous-valued function learning have also demonstrated clear limitations of exemplar models in situations involving continuous-valued function extrapolation (e.g., DeLosh, Busemeyer, & McDaniel, 1997; Juslin, Olsson, & Olsson, 2003). Apparently, more abstract forms of quantitative reasoning than can be achieved by exemplar-based generalization operate in such situations.

The present research differs from these previous lines of work in two major ways. First, as noted above, the previous paradigms generally involved situations in which observers were provided with continuous-valued feedback associated with individual training exemplars. Furthermore, the observers needed to learn continuous-valued functional relations to perform the tasks. By contrast, in the version of the probabilistic inference task under current consideration, observers are provided with only discrete, categorical feedback concerning which member of each training pair has a greater value on the criterion variable. Observers are never provided with information concerning the precise, continuous-valued magnitude of the criterion variable, so abstract forms of quantitative reasoning and continuous-function estimation are not required. Indeed, as we discuss in the section The Formal Models, under these training conditions the multi-attribute inference task comes extremely close to standard categorization tasks in spirit and structure, so there is a natural bridge for applying exemplar models in this domain.[1]

Second, despite providing only discrete pairwise feedback, we developed versions of the probabilistic inference task that provide strong qualitative contrasts between the predictions of decision-making algorithms such as TTB and WADD and those of exemplar models. Whereas in most previous designs, individual attributes tended to be highly diagnostic of the magnitude of the criterion variable, our designs require observers to be sensitive to interactions between the attribute values. Thus, as explained in the sections The Formal Models and Experiment 1, learning should be extremely difficult or impossible according to standard versions of TTB and WADD, but quite natural and easy according to the alternative exemplar models.

To preview, although we expected the results to provide a knock-out blow to TTB and WADD, we were surprised to discover that our tasks were indeed extraordinarily difficult for most subjects to learn (Experiments 1 and 2). Therefore, we followed the initial experiments with a study that involved extensive training (up to 10 days) on a paired-comparison design that required sensitivity to attribute interactions. In this case, subjects eventually learned the task. However, analyses of subjects' response times for making their paired-comparison decisions suggested that most did not use an exemplar-based strategy. Instead, it appeared that subjects recoded the attribute values, with certain cue configurations being redefined as emergent cues. Subjects then made their paired-comparison decisions by applying a set of hierarchically organized rules based on the priority of the recoded attribute values. This process is very much in the spirit of a generalized version of the TTB algorithm that operates on both single cues and configural ones. Taken together, the results provide strong challenges to exemplar-based accounts of performance in these correlated-cue tasks.

The Formal Models

We describe the formal models with respect to their application to the paradigm illustrated in Table 1, which is based on a pairwise learning task conducted by Lee and Cummins (2004). In this example, there are 16 items in the training set that vary along six binary-valued attributes and that vary continuously along a criterion variable. Values of 1 along each of the binary attributes tend

[1] In the context of a set of multi-attribute inference tasks, Juslin, Jones, et al. (2003) included a training condition with only dichotomous feedback and found evidence consistent with the use of exemplar-based inference. However, this condition involved standard categorization training, not the type of paired-comparison inference training that is the focus of the present work.

Table 1
Training Stimulus Patterns Used in Lee and Cummins (2004) and Bergert and Nosofsky (2007; Experiments 1 and 2)

Stimulus number	Stimulus pattern	Criterion value
1	0 0 0 1 0 0	16
2	0 1 0 0 1 0	18
3	0 0 1 0 0 1	21
4	0 0 0 1 1 0	25
5	0 0 0 0 1 0	31
6	1 0 0 0 1 1	40
7	0 0 1 1 1 1	44
8	1 1 0 1 0 0	51
9	1 1 1 0 0 1	62
10	1 1 0 0 1 0	70
11	1 1 0 1 1 1	97
12	1 1 1 1 0 0	104
13	1 1 1 1 1 1	280
14	1 1 1 1 0 1	285
15	1 1 1 0 1 0	347
16	1 1 1 1 1 0	444

Note. Cue validities for Attributes 1–6 are .968, .918, .815, .641, .565, and .547.

to indicate higher values of the criterion variable, and values of 0 tend to indicate lower values. However, each attribute provides only probabilistic information related to the criterion. As described in depth by Lee and Cummins and by Czerlinski, Gigerenzer, and Goldstein (1999), the structure illustrated in Table 1 is based on a real-world environment relating various geographic features of the Galapagos Islands to the criterion variable of number of different species on each of the islands. Because of the stimulus domain's naturalistic structure, researchers have suggested that adaptive algorithms such as TTB might fare quite well in accounting for people's judgments in this situation (see, e.g., Gigerenzer & Goldstein, 1996, for extensive discussion along these lines).

In Lee and Cummins's (2004) paradigm, on each trial, a pair of the multi-attribute alternatives was presented and the observer judged which member of the pair was greater on the criterion. Discrete feedback was then provided concerning the correct forced-choice alternative. Following this training, various test pairs were presented to evaluate the type of decision-making strategy that observers may have adopted for performing the task (see Table 2). In these critical test pairs, a single attribute with high diagnosticity pointed to Alternative A, whereas most remaining attributes pointed to Alternative B. Thus, the TTB and WADD models tended to make opposite predictions for such pairs.

To provide faithful tests of the original formulations of the models, Lee and Cummins (2004) made strong assumptions concerning the weights that observers assigned to each attribute. In particular, the weights were set at values related to the *cue validities* of the attribute values. The cue validity of an attribute is computed by dividing the number of correct pairwise discriminations afforded by a cue by the number of total discriminations (correct and incorrect) afforded by that cue. To illustrate, as shown in Table 1, for the present stimulus structure Attribute 1 has a high cue validity, whereas the cue validity for Attribute 6 is extremely low (close to .50). In addition to assuming fixed attribute weights related to the cue validities, the models tested by Lee and Cummins incorporated the assumption that observers would respond deterministically on the basis of the weighted evidence of the cues.

Bergert and Nosofsky (2007) considered generalized versions of the TTB and WADD models that relaxed these assumptions. In these generalized models, rather than fixing the values of the attribute weights a priori, the attribute weights were allowed to be free parameters. This generalization seemed reasonable because observers may not always be able to directly estimate the cue validities of the attribute values on the basis of their experience with the training pairs (e.g., Newell, 2005; Newell & Shanks, 2003). Furthermore, Bergert and Nosofsky made allowance for probabilistic forms of responding within the framework of the models. With respect to TTB, Bergert and Nosofsky's generalization allowed for the possibility that instead of using a fixed order of attribute inspection on each and every trial, observers might inspect attributes in a probabilistic order across trials. Likewise, in the strong version of the WADD model considered by Lee and Cummins (2004), the observer was assumed to respond deterministically with the alternative with the greater amount of summed evidence, even if the difference in summed evidence between the alternatives were minuscule. In the generalized model considered by Bergert and Nosofsky, the probability of choosing an alternative grew larger as the evidence difference between the alternatives grew larger. In experiments that followed up on the original studies of Lee and Cummins, Bergert and Nosofsky obtained clear evidence of the utility of the generalized models. These generalized versions of TTB and WADD serve as useful sources of comparison to the exemplar models considered in the present work.

Generalized TTB Model

The generalized TTB (gTTB) model is formally identical to Tversky's (1972) classic elimination-by-aspects model as applied to paired comparisons. According to the gTTB model, the observer assigns a weight w_m ($0 \leq w_m \leq 1$) to each individual attribute m. Without loss of generality, the weights are constrained to sum to 1. At any step of the process, the probability that the observer selects attribute m from a set of attributes S as a basis for making a decision is given by

$$P(m|S) = \frac{w_m}{\sum\limits_{k \in S} w_k}, \tag{1}$$

where S denotes the set of all attributes that have not yet been considered. If the selected attribute discriminates between the pair of alternatives (that is, if the binary-valued attribute has mismatching values across the pair of alternatives), then the observer

Table 2
Transfer Stimulus Patterns Used in Lee and Cummins (2004)

Transfer stimulus pair	TTB item	WADD item
1	1 0 0 0 0 1	0 1 1 0 0 0
2	1 0 0 0 1 0	0 1 1 0 0 0
3	1 0 0 0 1 1	0 1 1 1 0 0
4	1 0 0 1 1 0	0 1 1 1 0 0
5	1 0 0 1 1 1	0 1 1 1 1 0

Note. TTB item = item favored by the strong version of the take-the-best model; WADD item = item favored by the strong version of the weighted-additive model.

chooses the alternative with the positive attribute value. Otherwise, the attribute is removed from consideration, thereby forming a reduced set. A new attribute is then selected from the reduced set in accord with Equation 1. The process continues until a choice is made. As noted by Tversky (1972), the model predicts that the probability that Alternative A will be chosen over Alternative B is given by

$$P(A;A,B) = \frac{\sum_{a \in FA} w_a}{\sum_{a \in FA} w_a + \sum_{b \in FB} w_b}, \tag{2}$$

where FA denotes the set of all attributes that favor Alternative A, and likewise for FB.[2]

In addition, to provide a rudimentary form of error theory common to all of the compared models, Bergert and Nosofsky (2007) also made allowance for a guessing parameter g ($0 \leq g \leq 1$). The overall probability that a observer will choose Alternative A from pair (A,B) is given by $g/2 + (1 - g)TTB_A$, where TTB_A denotes the probability of selecting Alternative A according to the TTB process described above. Thus, in the present situation, the gTTB model has six free parameters: 5 freely varying attribute weights (w_m) and the guessing parameter g.[3]

WADD Model

In the version of the WADD model considered by Lee and Cummins (2004), the observer was assumed to sum the weights associated with all attributes that favored Alternative A and likewise for the weights associated with Alternative B. The observer then selected the alternative with the greater summed evidence. As described above, the weights were held fixed at values related to the cue validities of the attributes. In the generalization considered by Bergert and Nosofsky (2007), the probability that Alternative A was chosen over Alternative B was given by

$$P(A;A,B) = \frac{g}{2} + (1 - g)\frac{\left(\sum_{a \in FA} w_a\right)^{\gamma}}{\left(\sum_{a \in FA} w_a\right)^{\gamma} + \left(\sum_{b \in FB} w_b\right)^{\gamma}}, \tag{3}$$

where g is the guessing parameter, γ is a freely estimated response-scaling parameter (Ashby & Maddox, 1993; Nosofsky & Zaki, 2002), and the weights (w_m) are treated as free parameters, as in the gTTB model. Note that when $\gamma = \infty$ and $g = 0$ in Equation 3, the observer always chooses the alternative with the greater summed weighted evidence, so the present model reduces to Lee and Cummins's version of WADD as a special case. It is interesting that in quantitative tests of the WADD model, Bergert and Nosofsky (2007) found that a version with $\gamma = 1$ fitted the individual subject data quite well. Note that the $\gamma = 1$ version of WADD is formally identical to the gTTB model with respect to predicting choice probabilities (compare Equations 2 and 3). Thus, although the models are based on markedly different conceptual underpinnings, under certain parameter settings they nevertheless yield formally identical predictions.

Exemplar Models

The aim of the present research was to consider the application of exemplar-based models to the paired-comparison probabilistic inference task. We take as a starting point the generalized context model (GCM) of categorization (Nosofsky, 1986). It is straightforward to adapt the model to the present paradigm. We consider two such adaptations in the present article.

Model 1. Recall that in the present paradigm, on each training trial a pair of alternatives is presented and the observer receives feedback concerning the alternative that has a greater value on the criterion variable. In Model 1, we assume that the winning alternative is stored as an exemplar of the winners category, and the losing alternative is stored as an exemplar of the losers category. This same storage process is repeated for each individual training trial.

At time of transfer, a test pair (A,B) is presented. As is the case in the GCM, the observer is assumed to sum the similarity of Alternative A to all of the exemplars of the winners category ($S[A,W]$) and to all of the exemplars of the losers category ($S[A,L]$):

$$S(A,W) = \sum_{w \in W} s(A,w)$$

$$S(A,L) = \sum_{l \in L} s(A,l), \tag{4}$$

where $S(A,W)$ and $S(A,L)$ denote the similarities of A to each winning and losing exemplar, respectively. The relative goodness of Alternative A is then given by

$$G_A = \frac{S(A,W)}{S(A,W) + S(A,L)}. \tag{5}$$

An analogous process occurs for Alternative B. The probability that Alternative A is then judged as higher on the criterion variable than is Alternative B is given by

$$P(A;A,B) = \frac{g}{2} + (1 - g)\frac{(G_A)^{\gamma}}{(G_A)^{\gamma} + (G_B)^{\gamma}}, \tag{6}$$

where γ ($0 \leq \gamma$) is a response-scaling parameter.

The similarity of Alternative A to each stored exemplar is computed as in the GCM. Specifically, the distance between Alternative A and Exemplar j is given by

$$d(A,j) = \sum_{m} w_m |x_{Am} - x_{jm}|, \tag{7}$$

[2] Note that although this version of gTTB allows for probabilistic orders of attribute inspection, in situations in which the weights are highly noncompensatory the observer would tend to inspect the attributes in what is essentially a fixed, deterministic order. For example, if the magnitude of w_1 far exceeds the magnitude of w_2, the magnitude of w_2 far exceeds w_3, and so forth, then the observer would inspect the attributes in the order 1, 2, 3, 4, 5, 6. This fixed, deterministic order of inspection is assumed in applications of strong versions of the TTB model.

[3] Because the weights can be constrained to sum to one without loss of generality, the number of freely varying weights is one less than the number of attributes.

where x_{Am} and x_{jm} are the (binary) values of A and j on attribute m and where w_m ($0 \leq w_m \leq 1$, $\sum_m w_m = 1$) is the weight given to attribute m in computing distance. These freely estimated weights play an analogous role in the exemplar model as they do in the gTTB and WADD models discussed previously. Finally, the similarity between Alternative A and exemplar j is given by

$$s(A,j) = e^{-cd(A,j)}, \tag{8}$$

where c is a freely estimated sensitivity parameter (Shepard, 1987). These computed similarities are substituted into Equations 4–6 to generate the quantitative predictions from the model. This version of the exemplar model uses eight free parameters: the sensitivity parameter c, five freely varying attribute weights w_m, the response-scaling parameter γ, and the guessing parameter g.

In general, the model predicts that alternatives will be chosen if they are similar to previous exemplars that have been trained as winners, that is, as being high on the criterion variable. Furthermore, if the observer learns that some attributes are more important than are others in predicting the criterion, then such knowledge would be reflected by high weighting of those attributes.

Model 2. Whereas the previous model supposes that winners and losers are recorded separately in memory, another idea is that the observer may be sensitive to the precise pairs of alternatives that are experienced during training (Cohen & Nosofsky, 2000). Thus, in Model 2, if the observer receives feedback that Alternative A is higher on the criterion than is Alternative B, then a concatenated A–B representation is stored as a member of the winners category, and a concatenated B–A representation is stored as an exemplar of the losers category. In the present paradigm, each individual alternative is composed of six attribute values. Thus, each concatenated A–B representation is composed of 12 attribute values, the first 6 corresponding to Alternative A and the second 6 corresponding to Alternative B.

When a test pair is presented, one computes the similarity of the entire pair of items to each exemplar pair stored in memory, using equations that are analogous to those already given in Equations 7 and 8. (In the modeling involving the concatenated pairs, the weight assigned to Attribute $m + 6$ is the same as the weight assigned to Attribute m.) The probability that Alternative A is judged to be higher on the criterion than Alternative B is given by

$$P(A;A,B) = \frac{g}{2} + (1 - g)\frac{S(AB,W)^{\gamma}}{S(AB,W)^{\gamma} + S(AB,L)^{\gamma}}, \tag{9}$$

where $S(AB,W)$ is the summed similarity of pair A–B to all the exemplar pairs of the winners category, $S(AB,L)$ is the summed similarity of pair A–B to all the exemplar pairs of the losers category, and γ is the response-scaling parameter. This version of the exemplar model uses the same eight free parameters as does Version 1. Cohen and Nosofsky (2000) applied a similar version of the exemplar model to the analysis of paired-comparison same–different judgments. Rieskamp and Otto (2006) considered a special case of this paired-comparison exemplar model as well, although they did not conduct the types of experimental tests that form the theme of the present investigation.

Preliminary Tests

In preliminary tests of the exemplar-modeling ideas, we fitted the models to the sets of probabilistic-inference data reported in Bergert and Nosofsky's (2007) replication and extension of Lee and Cummins's (2004) experiment. We used the previously obtained fits from the gTTB model (Bergert & Nosofsky, 2007) as a basis for comparison. (Recall that the best-fitting version of the WADD model was formally identical to the gTTB model, so the WADD model is also part of these comparisons.) Because the nature of the experiments was described in detail by Bergert and Nosofsky, we provide only a brief summary here. The physical stimuli that instantiated the abstract structure of the stimulus set (Tables 1 and 2) were drawings of bugs that varied along six, binary-valued attributes (body parts) and a continuous-valued criterion (poisonousness). On each trial of the training phase, a pair of bugs was presented on the computer screen and the participant judged which member of the pair was more poisonous. Corrective feedback was provided following each training trial. The feedback indicated only which of the bugs was more poisonous; it did not indicate the continuous value of the criterion variable. There were two blocks of these training trials, with each of 119 unique pairs of bugs presented once each, in a random order, in each block.[4] Following training, there was a test phase that included a presentation of each of the training pairs, plus eight presentations of each of five test pairs that were useful for discriminating between the predictions of the strong versions of the TTB and WADD models. No feedback was presented during the test phase.

In the present analyses, we fitted both versions of the exemplar model to the test phase data of each of the individual subjects from Bergert and Nosofsky's (2007) experiment. Following Bergert and Nosofsky, we used the Bayesian Information Criterion (BIC; Schwarz, 1978) as a measure of fit.[5] The BIC penalizes a model for its number of free parameters: The BIC gets smaller as the absolute likelihood-based fit of a model improves but grows larger as the number of free parameters in a model increases. The model that yields a smaller BIC is considered to provide a better overall account of the data.

The results of these model-fitting analyses are presented in Table 3. The table shows, for each individual subject, the BIC fit yielded by the gTTB (and equivalent WADD) model and by each version of the exemplar model. This modeling analysis yields preliminary evidence in favor of the exemplar-based modeling ideas. Versions 1 and 2 of the exemplar model provided better BIC fits than did the gTTB (or WADD) model for 45 and 46 of the 61 subjects, respectively. In addition, the mean BICs yielded by the exemplar models (Version 1, $M = 104.4$; Version 2, $M = 103.3$) were both significantly smaller than the one yielded by the gTTB/WADD model ($M = 111.0$), mean $t(61) = 4.26$, $p < .001$. Thus, at the very least, the exemplar-based models appear to be highly competitive with two of the major contending models of multi-attribute, paired comparison inference, so are clearly worthy of further investigation.

[4] Pair 2–7 was deleted from training because the strong versions of the TTB and WADD models made opposite predictions for this pair. It was desired to establish an environment in which use of TTB and WADD strategies would make identical decisions on each of the training pairs.

[5] BIC $= -2*\ln(L) + p*\ln(N)$, where $\ln(L)$ is the maximum log-likelihood of the data given the parameters in the model, p is the number of free parameters, and N is the number of observations in the data set.

Table 3
BIC Fits of the Models to the Individual Subject Data From Bergert and Nosofsky (2006; Experiment 1)

Subject	gTTB	EX-1	EX-2
1[a]	65.2	68.9	62.5
2	189.2	175.3	173.2
3	82.4	74.2	72.2
4	155.3	149.0	143.6
5	191.0	182.5	185.2
6	135.9	89.9	85.9
7	77.3	71.1	66.1
8	145.8	128.4	129.9
9	133.7	123.7	126.5
10	94.6	93.6	94.0
11	106.4	90.9	90.0
12	173.3	161.1	161.1
13	77.6	65.0	51.8
14	88.1	83.2	75.3
15	125.3	110.6	109.9
16	112.8	87.1	86.5
17[ab]	42.9	61.3	68.8
18	82.0	78.7	78.7
19	107.2	101.0	99.6
20	80.4	72.8	71.0
21[ab]	78.1	108.7	81.9
22	110.4	87.8	87.8
23	102.4	93.5	98.6
24	97.8	84.7	75.3
25[ab]	60.3	64.7	61.8
26[ab]	67.0	83.2	82.5
27[ab]	199.6	210.2	210.4
28	67.9	66.1	54.5
29	66.9	66.5	66.8
30	135.0	126.1	124.5
31	52.0	47.4	43.8
32	112.6	93.4	91.5
33	70.6	64.4	64.6
34	102.7	78.3	85.9
35[a]	160.8	157.3	171.9
36	149.2	115.6	114.0
37[a]	95.3	99.2	99.8
38[b]	72.9	74.5	67.3
39	97.0	91.1	90.1
40	74.1	53.1	58.2
41[ab]	57.7	69.6	68.1
42	148.3	135.9	140.1
43	151.1	138.0	137.6
44[ab]	92.0	92.9	93.7
45	141.2	135.7	135.8
46	136.9	131.9	129.3
47	117.9	114.5	114.0
48	152.2	145.3	142.4
49	169.8	169.4	168.7
50[ab]	212.1	222.8	222.8
51[ab]	75.1	88.0	87.4
52[ab]	77.9	82.0	78.6
53	133.6	104.8	123.4
54[ab]	86.6	88.4	87.3
55[ab]	99.5	111.9	105.7
56	97.7	94.8	95.3
57[ab]	189.1	189.6	189.5
58	113.1	79.5	79.5
59	60.3	47.5	48.7
60	115.7	89.3	90.0
61	104.0	101.7	98.4
M	111.0	104.4	103.3

Note. BIC = Bayesian Information Criterion; gTTB = generalized take-the-best model; EX-1 = Version 1 of exemplar model; EX-2 = Version 2 of exemplar model.
[a] Case in which gTTB yields a better fit to the subject's data than does EX-1. [b] Case in which gTTB yields a better fit to the subject's data than does EX-2.

Although the preliminary fit indices favor the exemplar models, such fits need to be interpreted with caution. The BIC statistic penalizes a model for its increase in number of free parameters, but it is not sensitive to the inherent flexibility associated with models because of their functional form (Myung, 2000). That is, even if Model A has the same number of free parameters as does Model B, the functional form of Model A may still allow it to trace out more flexibly the universe of possible data patterns. One approach to dealing with this problem involves the use of more sophisticated model-fit statistics that penalize models for their intrinsic flexibility and complexity (Myung, Balasubramanian, & Pitt, 2000). In the present work, however, our strategy was to develop multi-attribute inference paradigms that would provide strong qualitative contrasts between the natural, a priori predictions from the models. As explained below, we then tested the ability of the models to generalize their predictions from the initial paradigm tested by Lee and Cummins (2004) to our new multi-attribute inference paradigms.

Experiment 1

The gTTB and WADD models can both be viewed as *independent-cue* models (e.g., Medin & Schaffer, 1978). In the gTTB model, the observer is presumed to inspect single cues, one at a time, until a discriminating cue is found. The decision is then based solely on the value of that single cue. Likewise, according to the WADD model, the observer sums evidence associated with values of individual cues. Both models are insensitive to information provided by combinations or interactions between cues. By contrast, the exemplar models are *interactive-cue* models and are highly sensitive to such information (for extensive previous discussions of this point, see, e.g., Medin & Schaffer, 1978).

As described in depth by Lee and Cummins (2004), the stimulus structure (Table 1) that these researchers tested was derived from a natural environment. Furthermore, in this environment, it happened to be the case that the individual cues provided roughly independent information concerning the magnitude of the criterion variable. Indeed, as noted by Lee and Cummins (2004), with appropriate settings of the various cue weights, observers could achieve 86% correct paired-comparison decisions with either the TTB or WADD strategies. It is straightforward to verify that again with appropriate settings of the parameter values, use of the exemplar-based strategies would allow observers to achieve perfect performance on the training set. The main point is that use of either an independent-cue or interactive-cue (e.g., exemplar-based) strategy would allow for reasonably good performance on the Table 1 structure.

To develop strong contrasting predictions from the models, in the present experiment we tested subjects by using the alternative correlated-cue structure shown in Table 4. This structure embeds two different biconditional rules for predicting the magnitude of the criterion variable. As can be seen from inspection of the table, values of 1–0 on Attributes 1 and 2 tend to predict the highest magnitudes; values of 0–1 tend to predict the next highest magnitudes; and values of 1–1 and 0–0 predict lower magnitudes. An analogous biconditional rule operates on Attributes 3 and 4 for predicting more fine-grained magnitudes of the criterion variable.

It is important for this structure that individual cues provide virtually no diagnostic information concerning the magnitude of

Table 4
Stimulus Patterns and Criterion Values Used in Experiment 1

Stimulus number	Stimulus pattern	Criterion value
1	0 0 0 0 0 0	0
2	0 0 1 1 0 0	4
3	1 1 0 0 0 0	15
4	1 1 1 1 0 0	19
5	0 0 0 1 0 0	21
6	0 0 1 0 0 0	25
7	1 1 0 1 0 0	36
8	1 1 1 0 0 0	40
9	0 1 0 0 0 0	60
10	0 1 1 1 0 0	64
11	1 0 0 0 0 0	75
12	1 0 1 1 0 0	79
13	0 1 0 1 0 0	81
14	0 1 1 0 0 0	85
15	1 0 0 1 0 0	96
16	1 0 1 0 0 0	100

Note. Cue validities for Attributes 1–4 are .625, .500, .563, and .500. Depending on the assumptions, validity is either undefined for Cues 5 and 6 because they make zero distinctions, or else validity is .500 if the decision process terminates with a guess.

the criterion variable (see individual cue validities listed in Table 4). We conducted a computer search for the magnitudes of the attribute weights that would maximize the percentage of correct paired-comparison decisions that gTTB or WADD (with free γ) could achieve for the task. The maximum performance that can be achieved by either independent-cue strategy is only 60% correct.

By contrast, as we will soon document, both of the exemplar-based strategies continue to predict very high levels of performance for this correlated-cues task. Regardless of whether cues combine in an independent or interactive manner, the exemplar model simply assesses the degree of similarity between the members of the tested pairs and the trained pairs and responds accordingly. For example, because an item with a 1–0 combination on Attributes 1 and 2 is highly similar to numerous exemplars from the trained winners category, there would be a great deal of evidence that this item is high on the criterion variable.

Our manner of contrasting the alternative models is therefore straightforward. We simply trained subjects on the correlated-cue structure, using the same methods Lee and Cummins (2004) and Bergert and Nosofsky (2007) used for the previous independent-cue structure, and then assessed subjects' performance on the subsequent test. The gTTB and WADD models predict extremely poor performance, whereas the exemplar models predict excellent performance. We develop the precise predictions from the formal models in the Results section of the experiment.

Method

Subjects. The subjects were 98 undergraduate students at Indiana University who participated to partially fulfill a requirement for an introductory psychology class. Subjects were told at the beginning of the experiment that if they performed well on the test phase of the experiment they would be paid a $3 bonus; those who achieved 80% correct or better were paid the bonus.

Stimuli. Table 4 contains the 16 abstract stimulus patterns used in Experiment 1 and their corresponding poison levels. Each stimulus's poison level was calculated as a sum of the contribution of the first two features and a contribution of the middle two features. The first two features of a stimulus contributed to its poison level using the rule: 1–0 = 75, 0–1 = 60, 1–1 = 15, and 0–0 = 0. The second two features of a stimulus contributed to its poison level using the rule: 1–0 = 25, 0–1 = 21, 1–1 = 4, and 0–0 = 0. The last two features were constant for all stimuli for each participant (the constant values were randomly determined). So, for example, the pattern 101100 gets a contribution of 75 from the first two features, a contribution of 4 from the middle two features, and a contribution of 0 from the last two features, for a total poison level of 79, which is the poison level listed in Table 4 for Stimulus 12.

These abstract stimulus patterns were mapped onto pictures of bugs (see Bergert & Nosofsky, 2007, for illustrations), with the six features corresponding to six body parts: body, head, legs, antennae, fangs, and tail. The binary values of these features were different appearances of the same feature (e.g., long or short legs).

The mapping of the six abstract features onto the six physical features was randomized across subjects, so that, for example, Feature 1 corresponded to the eyes for one participant and the legs for another subject. Similarly, the mappings of abstract values to physical feature levels were also randomized, so that, for example, long legs might indicate poison for some subjects, whereas short legs might indicate poison for other subjects.

Procedure. The experiment consisted of two blocks of training followed by one block of testing. Each block contained a two-alternative decision trial for each of the 120 possible pairs of the 16 stimuli. The order of the 120 pairs was randomized in each block for every subject.

During each trial of the training phase, a participant was shown a pair of bugs on a computer screen, one on the left and one on the right, and was asked to decide which of the two was more poisonous. The left–right positioning of the bugs in each pair was randomized on each trial. A participant gave his or her response by pressing the *F* key (labeled *LEFT*) or the *J* key (labeled *RIGHT*) on a computer keyboard. After responding, the participant was given feedback about which response was correct, in the form of a red rectangular border appearing around the more poisonous bug. The word *Correct* or *Incorrect* also appeared below the pair of bugs during this feedback. The participant was allowed to study the feedback, and the pair of bugs, for as long as he or she wished. To move on to the next trial, the participant pressed the space bar, making each trial self-paced.

The test phase differed from the training phase in that no feedback was provided; instead, following Lee and Cummins's (2004) procedure, subjects provided a confidence rating following each paired-comparison decision. (However, we do not analyze the confidence ratings in the present article.) As was the case in training, subjects initiated the next trial by pressing the space bar.

Subjects were given instructions before training that outlined the task, mentioned the reward for good performance, and pointed out and named the six features differentiating the bugs. Additional instructions provided before the test phase mentioned that feedback would no longer be provided but that confidence ratings would be collected.

Results

The percentage of correct paired-comparison decisions achieved by the individual subjects during the test phase is displayed in the histogram shown in Figure 1 (Panel A). It is immediately apparent that performance was extremely poor. The lion's share of the subjects (80 of 98) scored below 65% correct. Averaged across all subjects, the mean percentage of correct decisions was only 56.7. Most subjects are clustered very close to the chance performance level of 50% correct. This exceedingly poor performance is not the

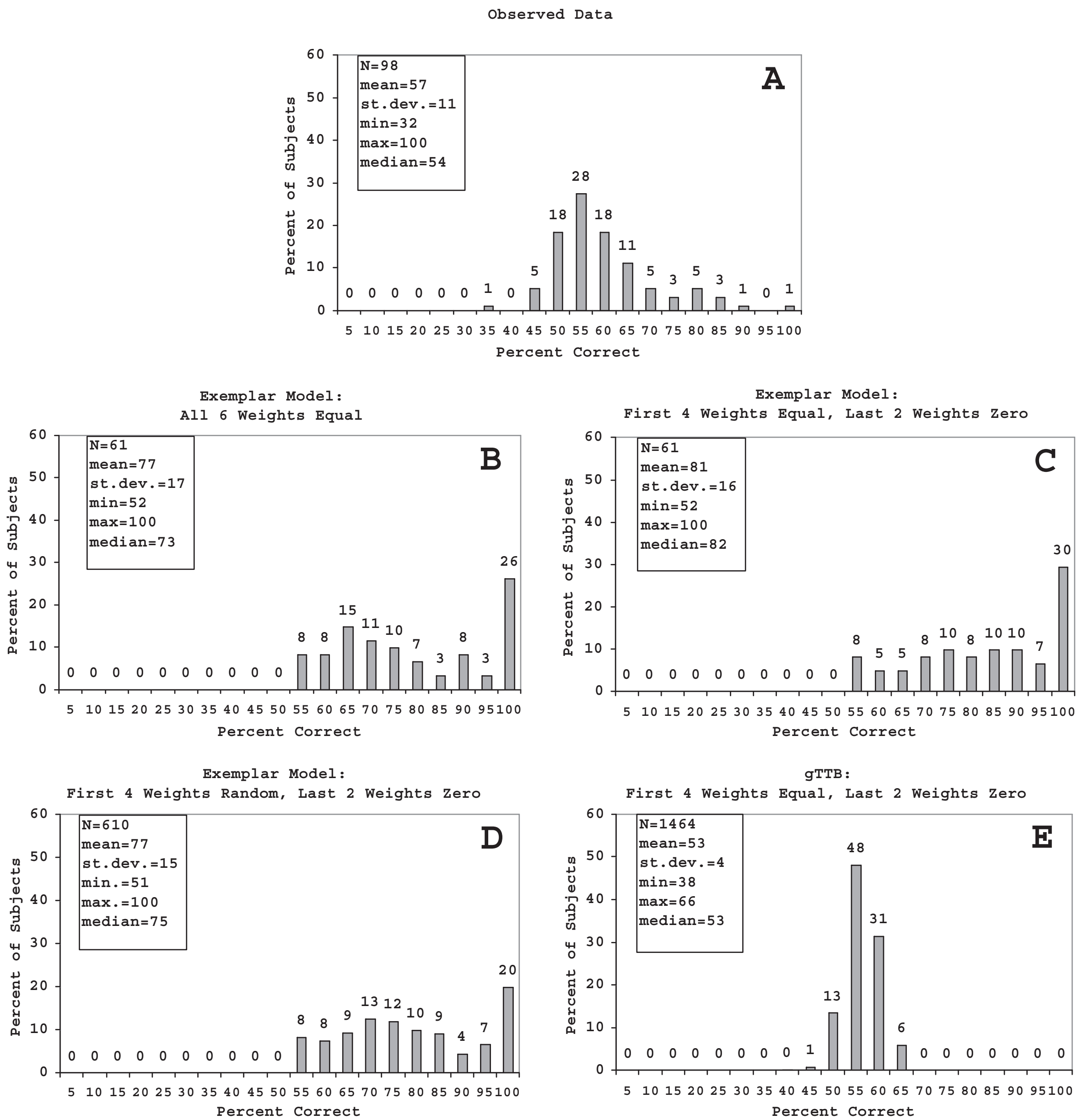

Figure 1. Observed and model-predicted performance in Experiment 1. The x-axis values indicate the upper limit of each class interval. A: The observed performance distribution of the 98 subjects. B: The performance distribution predicted by the exemplar model (Version 2) with all 6 weights set to be equal and all other parameters taken from fitting the model to the 61 subjects in Bergert and Nosofsky (2007). C: The exemplar model's predictions with the first four weights set to be equal and the last two weights set to zero. D: The exemplar model's predictions with the first four weights set to random values (10 random weight assignments per subject) and the last two weights set to zero. E: The predictions of generalized take-the-best (gTTB) and weighted-additive models with the first four weights set to be equal and the last two weights set to zero. N = 1464 because the 61 performance parameters estimated from the preliminary experiment are combined factorially with each of the 24 equally likely orderings of cue inspections in the simulations. St. dev. = standard deviation.

result of some computer programming error. The experimenters, who had explicit knowledge of the stimulus structure, achieved 100% correct performance. Likewise, as can be seen in the histogram, a few exceptional subjects were also able to achieve close to perfect performance in the task.

In our view, these results pose a significant challenge to the exemplar-based accounts of multi-attribute paired-comparison inference. Intuitively, if observers are basing their criterion judgments on the similarity of test items to exemplars stored in memory, then performance should be quite good.

To corroborate this intuition, we conducted a series of analyses to document the predictions from the exemplar models. Note that either version of the exemplar model can roughly fit the poor performance simply by setting the sensitivity or response-scaling parameters (Equations 6 and 8) at near-zero values or by setting the guessing parameter (Equation 6) at a value near unity. The interesting question, however, is what the exemplar models predict a priori.

To investigate this question, we used a form of generalization methodology advocated by Busemeyer and Wang (2000) as a basis for evaluating model performance. Specifically, we used the results from the exemplar-model fits to the preliminary data (Table 3) as a basis for predicting the overall performance levels for the present experiment. For each individual participant from the preliminary experiment, we held fixed the best-fitting values of the sensitivity, response scaling, and guessing parameters and then computed the exemplar-model predictions of overall performance in the present task. This method seems well justified, because the experiments used the same stimuli, same number of training trials, same training procedures, and subjects sampled from the same populations. All that differed across the experiments was the stimulus structure relating attribute values to the criterion. Because Versions 1 and 2 of the exemplar model yielded essentially equivalent predictions, for simplicity we report the results from only Version 2.

In the first analysis, we assumed that subjects weighted equally the six attributes that composed the stimuli. The distribution of predicted performance from this equal-weight model is shown in Panel B of Figure 1. As expected, the predicted performance is in a completely different ballpark from the observed data. The model predicts that subjects should perform quite well in the task.

Note that for the Table 4 structure, the values along Attributes 5 and 6 are constant across the stimuli. Because the values do not vary, it is likely that most observers learned to ignore them. Thus, rather than assuming equal weighting of the six attributes, it is perhaps more appropriate to assign zero weight to Attributes 5 and 6 and to assume an equal weighting of Attributes 1–4. The predictions from this equal-weight attribute-subset model are shown in Panel C of Figure 1. Predicted performance is now even more out of line with the observed performance in the task.

In another application of the exemplar model, we assumed that subjects distributed attention randomly across the four relevant attributes. For each participant from the preliminary experiment, we again held fixed the values of the sensitivity, response-scaling, and guessing parameters. Now, however, for each subject we constructed 10 different random weight vectors (i.e., we created 10 randomly simulated subjects from each of the original subjects). For each of these simulated subjects, we computed the predicted overall performance from the exemplar model. The distribution of predicted performance from this random-weight model is shown in Panel D of Figure 1. Once again, the predicted distribution of performance is in a different ballpark from the observed data.

As a source of comparison, we also generated predictions of individual subject performance from the gTTB (and equivalent WADD) model. For each subject from the preliminary experiment, we held fixed the estimated value of the guessing parameter. We then simulated the predicted performance from gTTB, assuming that subjects weighted equally the four varying attributes. (Only slightly higher performance would be predicted if subjects always inspected the cues in their order of cue validity.) Panel E of Figure 1 shows that the gTTB model's predictions are in the same range as the vast majority of subjects' observed performances in this correlated-cues task.

It is important to note, however, that 31% of the subjects did achieve performance levels greater than the maximum (60% correct) that could be achieved by using an independent-cue strategy. Apparently, therefore, a subset of the subjects showed at least some sensitivity to the correlated-attribute structure, reinforcing previous findings that different clusters of individual differences are often observed in tasks of classification and inference learning (e.g., Erickson, 1999; Lee & Webb, 2005; Nosofsky, Palmeri, & McKinley, 1994). In the main, however, performance hovered in the vicinity of what is predicted by a TTB or WADD strategy and was far worse than what is predicted by an exemplar-based strategy.

Model-Fitting Analyses

The preceding analyses of overall performance cast strong doubt on the predictions from the exemplar model. To investigate the results further, however, we conducted quantitative modeling analyses in which the formal models were fitted to the individual subject data with all parameters allowed to vary freely. We introduced two minor modifications to the gTTB model to allow reasonable applications to the present correlated-cues paradigm. First, note that for the present paradigm, the assignment of which cue value is the positive cue value is nearly arbitrary, because the cue validities are close to .50. Thus, the fit procedure made allowance for the model to decide, for each individual subject, the direction of choice for each cue value. Second, the gTTB model tested here assumed a slightly different guessing process than was formalized in Bergert and Nosofsky's (2007) study. The modification is described in more detail in Appendix A.

The quantitative fits of the exemplar and gTTB models to the individual subject data are reported in Table 5. As was the case in our preliminary tests (Table 3), we again used the BIC statistic as the criterion of fit. Examination of the table reveals an extremely clear pattern of results. The exemplar model yielded a better fit than did the gTTB model to the data of all 13 observers whose performance exceeded 70% correct in the test phase. (Recall that the maximum performance that can be achieved by a TTB or WADD strategy for the present paradigm is 60% correct.) By contrast, the gTTB model yielded better quantitative fits than did the exemplar model to the data of 80 of 85 observers whose overall performance was less than 70% correct. Thus, the results from the quantitative model-fitting analyses converge with the results ob-

Table 5
BIC Fits of the Models to the Individual Subject Data from Experiment 1

Subject	gTTB	EX-2	P(C)	Subject	gTTB	EX-2	P(C)
*60	189.8	204.0	0.32	89	164.0	199.9	0.54
*8	137.7	204.7	0.42	*5	187.0	204.7	0.55
*82	144.7	196.9	0.42	*7	195.1	202.4	0.55
*74	181.1	202.1	0.43	*18	176.2	201.9	0.55
*87	190.1	204.7	0.43	*41	162.8	201.3	0.55
*76	130.7	204.7	0.44	*59	154.8	177.8	0.55
*23	171.4	185.0	0.45	*72	122.5	160.2	0.55
*79	126.2	204.7	0.45	*81	93.7	159.7	0.55
*12	151.0	200.6	0.46	*19	177.2	181.9	0.56
*20	183.2	204.7	0.47	*47	96.0	167.9	0.56
*52	152.5	204.7	0.47	*57	132.9	158.1	0.56
*75	195.1	204.7	0.47	*91	135.0	197.7	0.56
*11	158.2	203.6	0.48	*95	65.1	116.8	0.56
*16	178.7	200.4	0.48	*62	149.2	165.1	0.57
*33	153.7	203.1	0.48	*37	109.6	136.1	0.58
*34	156.2	196.4	0.48	*61	195.1	203.5	0.58
*42	185.4	201.8	0.48	*73	125.0	188.4	0.58
*55	169.3	200.6	0.48	*94	92.7	158.9	0.58
*78	97.9	204.7	0.48	*32	115.5	150.5	0.59
*80	185.6	202.5	0.48	*3	171.0	186.9	0.60
*90	195.1	202.8	0.48	67	163.7	162.6	0.60
*97	138.1	185.1	0.48	*92	187.3	197.2	0.60
*36	123.1	185.6	0.49	*96	75.6	146.4	0.60
*64	192.8	204.7	0.49	*63	130.8	170.1	0.61
*1	142.0	173.9	0.50	70	142.3	121.3	0.61
*2	195.1	203.0	0.50	22	181.5	159.4	0.62
*6	157.0	193.0	0.50	*46	125.7	134.2	0.62
*26	185.9	202.4	0.50	*68	180.5	188.8	0.62
*54	104.5	116.3	0.50	*24	181.7	195.5	0.63
*58	192.3	202.9	0.50	*93	170.7	182.9	0.63
*65	149.0	185.6	0.50	13	195.1	185.5	0.65
*86	131.5	176.0	0.50	31	161.9	157.5	0.65
*25	195.1	203.8	0.51	*15	188.3	192.8	0.66
*44	146.7	188.4	0.51	*27	174.5	186.0	0.66
*71	181.6	189.8	0.51	*49	178.7	180.2	0.68
*10	182.1	202.2	0.52	39	195.1	144.3	0.72
*28	166.2	181.4	0.52	48	188.1	153.0	0.72
*30	159.5	191.0	0.52	35	128.2	93.0	0.74
*77	123.3	159.5	0.52	66	195.1	114.9	0.77
*88	115.3	139.0	0.52	83	177.2	112.3	0.77
*9	110.6	202.1	0.53	98	190.1	111.4	0.78
*17	179.6	188.9	0.53	38	185.5	116.0	0.79
*21	179.5	199.5	0.53	85	179.6	91.5	0.79
*29	168.5	179.6	0.53	56	173.5	103.3	0.81
*43	174.7	199.7	0.53	4	171.4	123.9	0.82
*45	195.1	204.4	0.53	40	174.2	95.0	0.83
*50	179.9	203.6	0.53	53	191.0	99.5	0.87
*69	176.3	191.7	0.53	51	189.8	38.4	1.00
*84	146.1	202.5	0.53	*M*	159.3	174.9	0.57
*14	88.3	171.1	0.54				

Note. BIC = Bayesian Information Criterion; gTTB = generalized take-the-best model; EX-2 = Version 2 of exemplar model; P(C) = probability correct; * = denote cases in which gTTB yields a better fit to the subject's data than does EX-2.

tained by examining the overall performance profiles of the individual observers. For the vast majority of observers, performance is far more in line with the predictions from the independent-cue models than the exemplar model.

These quantitative model-fitting analyses are also important because they rule out an argument for saving the exemplar-based modeling ideas. In our previous analyses, we predicted the overall performance profiles (Figure 1) by assuming that observers would adopt a reasonable pattern of attribute weighting in which the relevant pairs of attributes were attended. One approach to saving the exemplar model would be to posit that observers attended to nondiagnostic configurations of attributes. For example, an observer might attend to Attribute 1 but fail to attend to Attribute 2. However, the present modeling analyses make allowance for even this highly suboptimal pattern of attribute weighting, and the exemplar model still performs worse than does the gTTB model at fitting the data. Thus, the results challenge the exemplar model even in its general form.

Experiment 2

Achieving good performance for the correlated-cue structure tested in Experiment 1 would require subjects to attend simultaneously to all four relevant attributes of the stimuli. Furthermore, it would require simultaneous sensitivity to two biconditional rules defined across two separate pairs of attributes. Previous tests of exemplar models in the domain of classification indicate that subjects do learn to attend to relevant attributes of the objects and that they can learn biconditional structures (e.g., Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). However, the process is sometimes a gradual one, and very early in learning observers focus on only subsets of relevant attributes (Nosofsky & Zaki, 2002).

Therefore, to follow up on the results from Experiment 1, in the present experiment we designed a new stimulus structure that made more moderate demands on subjects' sensitivity to correlated attributes. The structure is shown in Table 6. In this case, there is only a single biconditional rule, defined over Attributes 1 and 2, that is relevant to predicting values on the criterion variable. Cue combinations of 1–0 and 0–1 on Attributes 1 and 2 predict the highest values of the criterion variable, whereas cue combinations of 1–1 and 0–0 predict lower values. Attribute 1 also has moderate individual cue validity (.75), whereas Attribute 2 has individual cue validity equal to .50. Attributes 3 and 4 have relatively low individual cue validities (.625 and .5625, respectively), and their combinations are also not highly predictive of the criterion variable. The best that the TTB or WADD (free γ) strategies can achieve for the Table 6 structure is 73% correct paired-comparison decisions. By contrast, as we document below, the most natural prediction from the exemplar models is once again that observers will perform extremely well in the task. The basic question was whether or not the exemplar-model prediction would shine through in this more moderate case in which only a single biconditional rule needed to be mastered.

Table 6
Stimulus Patterns and Criterion Values Used in Experiments 2 and 3

Stimulus number	Stimulus pattern	Criterion value
1	0 0 0 0 0 0	0
2	0 0 0 1 0 0	1
3	0 0 1 0 0 0	10
4	0 0 1 1 0 0	11
5	1 1 0 0 0 0	15
6	1 1 0 1 0 0	16
7	1 1 1 0 0 0	25
8	1 1 1 1 0 0	26
9	0 1 0 0 0 0	60
10	0 1 0 1 0 0	61
11	0 1 1 0 0 0	70
12	0 1 1 1 0 0	71
13	1 0 0 0 0 0	75
14	1 0 0 1 0 0	76
15	1 0 1 0 0 0	85
16	1 0 1 1 0 0	86

Note. Cue validities for Attributes 1–4 are .750, .500, .625, and .563. Depending on the assumptions, validity is either undefined for Cues 5 and 6 because they make zero distinctions, or else validity is .500 if the decision process terminates with a guess.

Method

Subjects. The subjects were 100 undergraduates from Indiana University who participated in partial fulfillment of a requirement of an introductory psychology class. The same bonus for good performance was used as in Experiment 1.

Stimuli. Table 6 contains the 16 abstract stimulus patterns used in Experiment 2 and their corresponding poison levels. Each stimulus's poison level was calculated as a sum of the contribution of the first two features and the contribution of the middle two features. The first two features of a stimulus contribute to its poison level using the rule: 1–0 = 75, 0–1 = 60, 1–1 = 15, and 0–0 = 0. The third feature contributes 10 to the poison level: 1 = 10 and 0 = 0. The fourth feature contributes 1: 1 = 1 and 0 = 0. The last two features were constant for all stimuli for each subject. These abstract stimulus patterns were mapped onto pictures of bugs in the way previously described for Experiment 1.

Procedure. The procedure was identical to the one used in Experiment 1.

Results

The observed distribution of individual observer performances is shown in Figure 2 (Panel A). Mean performance (63% correct) is somewhat higher than in Experiment 1 but is still quite poor overall. The vast majority of subjects (76 of 100) had performance levels that varied between 45% and 75% correct.

The predicted distributions of performance from the exemplar model (Version 2) are shown in Panels B–D of Figure 2. The specific assumptions that were made in generating these predictions are the same as for the previous ones in Figure 1. Regardless of the precise assumptions, the exemplar model again predicts performance levels considerably higher than are seen in the observed data. The model predicts that the majority of subjects should be performing at levels of 80% or above, with a high proportion of subjects predicted to achieve nearly perfect performance.

By comparison, the predictions from the gTTB model are shown in Panel E of Figure 2. As was the case for the analysis in Experiment 1, these simulated predictions from the gTTB model assume that the four relevant attributes were weighted equally. The gTTB model predicts the bulk of the observed distribution of performance quite well, with most subjects predicted to lie in the 55%–75% range.

Although the vast majority of subjects show performance levels in the vicinity of what is predicted by gTTB, there is again evidence of a smaller subset of the subjects exceeding the limits imposed by the independent-cue strategies. Such subjects are showing some sensitivity to the correlated-cue structure along Attributes 1 and 2, a result that we investigate in depth in Experiment 3.

Model-Fitting Analyses

We again fitted the gTTB and exemplar models to the individual subject data with all parameters allowed to vary freely. Inspection of Table 7 reveals an equally clear picture as obtained previously (we discuss the model labeled configural-cue TTB [cc-TTB] after presenting the results from Experiment 3). The exemplar model achieved a better quantitative fit than did the gTTB model for all 21 observers whose overall performance exceeded 77% correct deci-

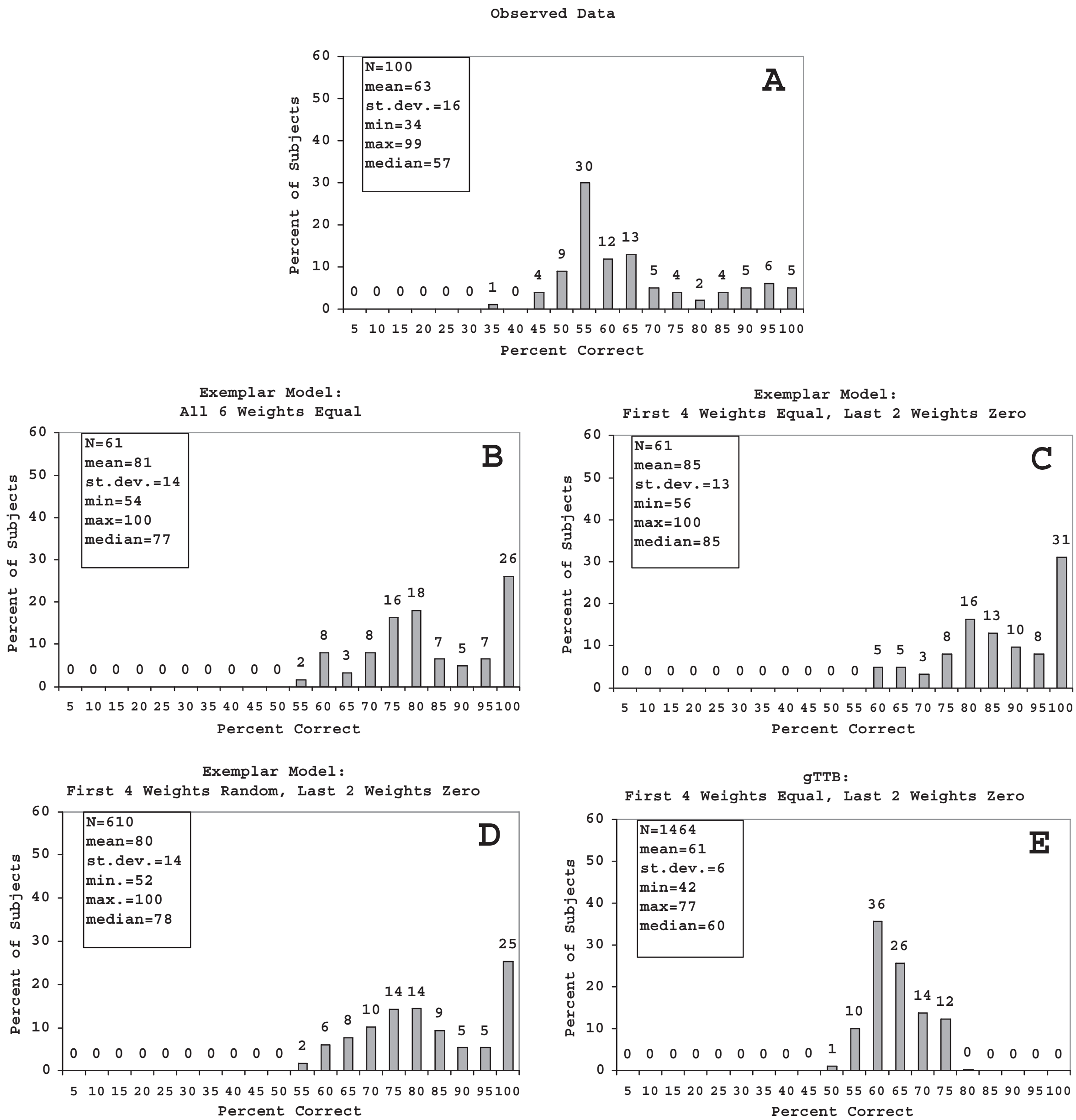

Figure 2. Observed and model-predicted performance in Experiment 2. The x-axis values indicate the upper limit of each class interval. A: The performance distribution of the 100 subjects. B: The distribution of performance predicted by the exemplar model (Version 2) with all six weights set to be equal and other parameters taken from the 61 subjects in Bergert and Nosofsky (2007). C: The exemplar model's predictions with the first four weights set to be equal and the last two weights set to zero. D: The exemplar model's predictions with the first four weights set to random values (10 random weight assignments per subject) and the last two weights set to zero. E: The predictions of generalized take-the-best (gTTB) and weighted-additive models with the first four weights set to be equal and the last two weights set to zero. St. dev. = standard deviation.

sions. By contrast, the gTTB model fitted better than did the exemplar model for all 79 observers whose overall performance was less than 77% correct. Thus, as was the case in Experiment 1, the results from both the overall performance profile analyses and the individual subject model fits point toward the independent-cue models and away from the exemplar model for the vast majority of subjects. A smaller subset of subjects, however, did show some sensitivity to the correlated attributes. In Experiment 3, we investigate further the possible bases for these good-performing subjects' behavior.

Experiment 3

We were surprised at the exceedingly poor performance exhibited by most subjects in our Experiments 1 and 2 and also inter-

Table 7
BIC Fits of the Models to the Individual Subject Data From Experiment 2

Subject	gTTB	EX-2	cc-TTB	P(C)	Subject	gTTB	EX-2	cc-TTB	P(C)
*54	184.6	204.1		0.34	*17	150.3	158.1		0.58
*49	195.1	204.7		0.42	*23	187.0	199.9		0.58
*62	187.9	199.3		0.43	*37	171.4	178.4		0.58
*97	177.8	198.5		0.43	*45	195.1	200.4		0.58
*64	161.7	199.5		0.44	*69	111.0	133.1		0.58
*86	186.7	204.1		0.45	*76	79.8	179.4		0.58
*98	191.9	204.2		0.47	*46	130.2	137.6		0.60
*73	157.8	204.2		0.48	*48	119.8	148.5		0.60
*7	154.3	192.0		0.49	*52	117.6	155.3		0.60
*31	149.6	192.0		0.49	*92	156.6	161.9		0.60
*47	186.6	203.2		0.49	*24	182.2	190.8		0.61
*63	162.9	175.3		0.49	*59	155.6	172.1		0.61
*67	195.1	201.4		0.49	*74	114.7	133.3		0.61
*80	145.8	184.7		0.49	*95	153.6	175.7		0.61
*5	195.1	204.4		0.50	*41	135.9	158.5		0.62
*14	143.8	176.0		0.50	*85	129.9	140.0		0.62
*21	177.3	203.4		0.50	*83	108.3	151.1		0.63
*40	190.6	204.6		0.50	*93	138.3	141.8		0.63
*57	189.7	196.4		0.50	*56	133.8	168.4		0.64
*71	165.1	180.0		0.50	*27	162.1	167.1		0.66
*38	161.2	198.0		0.51	*44	122.1	130.4		0.67
*65	187.1	201.4		0.51	*18	67.3	94.6		0.68
*26	191.7	201.6		0.52	*51	93.1	133.9		0.68
*84	152.5	158.7		0.52	*58	90.8	117.3		0.68
*1	107.8	187.1		0.53	*9	161.5	165.6		0.70
*3	181.9	193.5		0.53	*89	88.5	106.9		0.71
*10	116.6	159.6		0.53	*22	94.6	101.2		0.72
*12	179.1	196.8		0.53	*4	28.7	49.5		0.73
*13	173.3	183.6		0.53	*43	161.4	167.8		0.75
*60	167.5	183.0		0.53	68	168.4	162.5	155.5	0.77
*61	160.4	168.9		0.53	72	184.3	150.3	134.9	0.81
*66	195.1	203.8		0.53	34	152.8	115.4	122.3	0.82
*82	141.3	150.9		0.53	53	175.9	147.4	143.6	0.82
*87	91.9	133.2		0.53	55	161.9	136.8	129.2	0.82
*90	166.9	178.3		0.53	6	190.0	133.3	128.2	0.85
*94	172.0	196.9		0.53	8	189.9	127.9	114.1	0.85
*96	173.8	190.6		0.53	32	168.5	128.2	82.7	0.87
*20	153.4	165.8		0.54	25	182.0	93.5	77.7	0.89
*29	141.2	179.4		0.54	81	161.8	103.7	111.7	0.89
*70	189.4	200.7		0.54	11	168.6	85.5	78.2	0.92
*75	156.1	172.4		0.54	88	176.2	93.1	71.0	0.92
*77	186.6	199.6		0.54	100	168.6	85.5	78.2	0.92
*91	195.1	201.4		0.54	50	179.5	78.0	64.7	0.93
*99	188.2	198.3		0.54	78	170.1	85.2	70.0	0.93
*33	195.1	198.6		0.55	15	174.4	73.6	65.8	0.94
*42	195.1	197.7		0.55	2	172.5	71.7	51.9	0.96
*79	138.0	161.8		0.55	19	174.5	76.7	71.6	0.96
*16	195.1	199.4		0.57	30	171.2	63.0	58.7	0.97
*36	186.4	193.9		0.57	28	173.3	49.6	43.9	0.99
*39	126.6	185.6		0.57	35	172.3	49.3	37.3	0.99

Note. BIC = Bayesian Information Criterion; gTTB = generalized take-the-best model; EX-2 = Version 2 of exemplar model; cc-TTB = configural-cue take-the-best model; P(C) = probability correct; * = denote cases in which gTTB yields a better fit to the subject's data than does EX-2.

ested to discover what processes the small subset of good performers may have used to take advantage of the correlated attributes. A reasonable hypothesis is that observers start by adopting independent-cue strategies such as TTB or WADD. When they discover that such strategies fail to yield adequate performance, they shift to alternative strategies that are sensitive to the correlations among the attributes. The small subset of subjects who exhibited good performance in Experiments 1 and 2 may have initiated such a strategy shift. Conceivably, numerous other subjects may have followed suit with more extended training. In Experiment 3, we pursue this possibility by repeating the design tested in Experiment 2; however, now rather than using only a couple of training blocks, we tested a smaller number of subjects for up to 10 multiple-block sessions of training, with each session conducted on a separate day.

Furthermore, to gain more information about the kinds of strategy shifts that the extended training might induce, we recorded subjects' response times (RTs) for making their paired-comparison decisions. As we discuss in the Theoretical Analysis section, even if subjects learn to make their paired-comparison decisions with perfect accuracy, the exemplar model predicts differential RTs for the individual paired-comparisons. We also consider an extended version of the TTB model that assumes that subjects learn to recode the correlated attributes into emergent configural cues, but thereafter to make decisions in accord with a standard TTB process (cf. Garcia-Retamero, Hoffrage, Dieckmann, & Ramos, 2007). This model too predicts interesting patterns of RTs for different types of paired-comparison decisions. We then use the observed RT data to distinguish between the predictions of the exemplar model and the cc-TTB model.

Method

Subjects. The subjects were 11 members of the Indiana University community who were paid $8 per session. In addition, the same bonus for good performance was used as in Experiments 1 and 2. The bonus applied to each individual session.

Stimuli. The same 16 abstract stimulus patterns were used as in Experiment 2 (see Table 6). These abstract stimulus patterns were mapped onto pictures of bugs in the way previously described for Experiments 1 and 2.

Procedure. Individual subjects participated for 8–10 sessions of training instead of just a single session. Each training session was conducted on a separate day. During each session, there were four blocks of training trials. Each block contained all 120 pairs of stimuli presented in a random order, and feedback was provided on each trial. We no longer asked subjects to provide confidence judgments for their paired-comparison decisions.

Results

The mean percentage of correct paired-comparison decisions is reported for each subject in each session of training in Table 8.[6] For Session 1, the data closely resemble the pattern of results observed in Experiment 2. The majority of subjects show very poor performance, with their percentage correct falling below the 73% maximum allowed by an independent-cue TTB strategy, and far short of the performance levels predicted by an exemplar-based strategy. As is evident from inspection of the table, however, soon after Session 1, the performance of most of the subjects begins to skyrocket. By the final sessions, all subjects are showing nearly perfect accuracy on the task.

Table 8
Mean Percentage of Correct Paired-Comparison Decisions for the Individual Subjects as a Function of Sessions of Training in Experiment 3

	Session									
Subject	1	2	3	4	5	6	7	8	9	10
1	75	95	99	100	99	99	99	99	98	99
2	63	80	95	96	95	98	98	96		
3	54	64	91	100	97	97	98	95	97	97
4	58	64	71	79	87	88	93	97	97	96
5	84	80	79	85	85	90	94	96	97	97
6	80		89	99	100	99	99	99	99	99
7	51									
8	57	74	82	94	99	100	99	99	100	100
9	91	99	98	98	98	98	98	99	99	98
10	58	67	86	96	95	95	96	95		
11	53	71	91	97	98	98	99	100		

Theoretical Analysis

Given the extremely high accuracies, it would be superfluous to try to model the choice-probability data with the independent-cue TTB and WADD models, both of which predict maximum accuracies of only 73% correct. Therefore, we focus our analyses on the patterns of RT data. As demonstrated below, these data are best presented from the perspective of the models themselves. We start, therefore, by describing the cc-TTB model.

cc-TTB Model

Consideration of the abstract stimulus structure in Table 6 suggests a simple strategy for making paired-comparison decisions regarding the magnitude of the criterion variable. First, subjects can recode the individual values along Attributes 1 and 2 into one of four configural cues: 1–0, 0–1, 1–1, or 0–0. If the members of a stimulus pair mismatch along this configural cue, then an immediate paired-comparison decision can be made, without checking the remaining attribute values. In particular, cue 1–0 always leads to the highest values of the criterion variable, cue 0–1 leads to the next highest values, cue 1–1 follows after that, and cue 0–0 leads to the lowest values of the criterion variable. Next, if the members of a stimulus pair match on the configural cue but mismatch on Attribute 3, then the value of Attribute 3 allows an immediate paired-comparison decision to be made. If Stimulus A has value 1 on Attribute 3, whereas Stimulus B has value 0 on Attribute 3, then Stimulus A is higher on the criterion variable. Finally, if the members of a stimulus pair match on both the configural cue and on Attribute 3, then Attribute 4 determines the ordering: Stimuli with value 1 on Attribute 4 are higher on the criterion variable than are stimuli with value 0.

A natural hypothesis, therefore, is that subjects engage in a set of hierarchically organized rule-based decisions for making their paired comparisons, much in the spirit of a TTB process. That is, subjects first compare the stimuli on the configural cue. If there is a mismatch, then the subject chooses the stimulus with the higher level of the configural cue. By contrast, if the stimuli match on the configural cue, then the subject inspects Cue 3 and so forth until a mismatching cue value is found. As described above, we imagine that the cue-inspection process takes place in a more-or-less serial fashion, with the configural cue inspected first, followed by Cue 3 and finally Cue 4. We should emphasize, however, that there is no logical necessity of a strictly serial inspection process. For example, observers may be able to obtain information about cue matches and mismatches in parallel fashion as well (see further discussion below). We refer to the model as the cc-TTB model (for application of a similar model to choice-probability data, see Garcia-Retamero et al., 2007).

[6] Subject 7 was unable to continue the experiment after Session 1, but this subject's Session 1 result is reported as well. In addition, the Session 2 data from Subject 6 were inadvertently overwritten.

To formalize the model, we allowed free parameters for representing the time needed to determine the direction of choice for each pair of mismatching configural cues: $T_{10\text{-}01}$, $T_{10\text{-}11}$, $T_{10\text{-}00}$, $T_{01\text{-}11}$, $T_{01\text{-}00}$, and $T_{11\text{-}00}$. That is, T_{ij} is the time needed to determine the direction of choice for a pair in which one stimulus has configural-cue i and the second stimulus has configural-cue j. Likewise, we defined a free parameter T_3 for the case in which a pair of stimuli matches on the configural cue but mismatches on Attribute 3. Finally, we defined a free parameter T_4 for the case in which a pair of stimuli matches on both the configural cue and on Attribute 3 but mismatches on Attribute 4. For example, the RT for pair 101100–010100 would be given by $T_{10\text{-}01}$, the RT for pair 101100–100100 by T_3, and the RT for pair 101100–101000 by T_4. Thus, this cc-TTB model makes use of eight free parameters for predicting mean RTs for 120 paired-comparison decisions.

It should be noted that various complex subprocesses would be expected to influence the values of these free parameters, so it is difficult to predict their values a priori. For example, $T_{01\text{-}00}$ might be faster than $T_{11\text{-}00}$ because the distance on the criterion variable is greater for 01/00 comparisons than for 11/00 comparisons (see Table 6). On the other hand, it is likely easier to determine that there is a mismatch for 11/00 pairs than for 01/00 pairs (i.e., there is a greater number of mismatching individual attributes for the former than for the latter). These competing factors make it difficult to predict a priori the ordering of the relevant free parameters. In addition, as noted above, although there is a strict ordering of priority of the attribute-based rules (i.e., the configural cue has highest priority, followed by Attribute 3 and finally Attribute 4), it is unclear whether a strictly serial process underlies the application of these rules. For example, for the present perceptual displays, a observer may be able to determine in parallel that two stimuli match on the configural cue and on Attribute 3, and attention might then be directed quickly to Attribute 4. Finally, because individual physical attribute values vary in their intrinsic salience, the time to locate matches and mismatches in these perceptual displays vary as well. Thus, for the present design, we allow the parameters to vary freely.

In our main analysis, we tested the model by fitting it to the mean correct RT data obtained from all of the subjects for those sessions in which accuracy was at 95% correct or above. (In subsequent analyses, we verify that individual subjects showed similar patterns as observed in these group-averaged data.) These mean RT data are reported for the 120 stimulus pairs in Appendix B.

We fitted the model to the data by searching for the values of the free parameters that minimized the sum of squared deviations between the predicted and observed mean RTs. The results are shown in Figure 3. Note that the figure distinguishes between eight stimulus types defined by the free parameters allowed in the cc-TTB model. The first six types correspond to pairs in which there is a mismatch on the configural cue. We denote these six types 10/01, 10/11, 10/00, 01/11, 01/00, and 11/00. For the seventh type (denoted Cue 3), the pairs match on the configural cue but mismatch on Attribute 3. For the final type (denoted Cue 4), the pairs match on both the configural cue and on Attribute 3 but mismatch on Attribute 4. Basically, the model predicts that individual tokens of these main stimulus types should have the same mean RTs, with each stimulus type being associated with a distinct processing-time parameter. As can be seen in Figure 3, the model provides a good description of the mean RT data, accounting for 92.4% of the response variance.

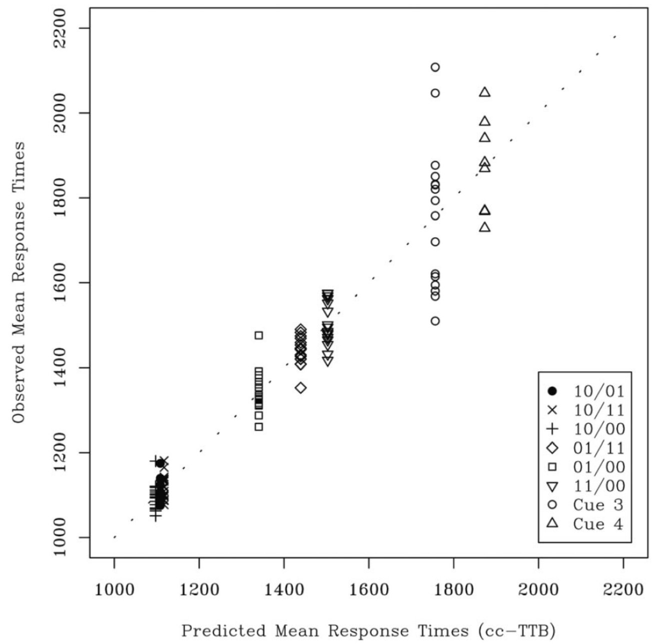

Figure 3. Experiment 3. Mean observed response times for the 120 stimulus pairs plotted against the predictions from the configural-cue take-the-best (cc-TTB) model.

Furthermore, inspection of the observed and predicted RTs reveals a highly interpretable pattern of results. The fastest RTs were observed for all pairs in which configural cue 1–0 was present on one member, whereas a mismatching configural cue was present on the other member. Note from Table 6 that configural cue 1–0 always wins in its paired comparisons. Intuitively, configural cue 1–0 is indicative of the killer bugs with the highest poison values. The presence of this cue on just one of the bugs allows subjects to make an immediate decision, without consideration of the details of the configural cue on the other member of the pair. By contrast, intuitively, the paired-comparisons for types 01/11, 01/00, and 11/00 require more care, because the stronger configural cue has not received consistent training as a winner; rather, its status depends on the value of the competing member of the pair.[7] Finally, pair types Cue 3 and Cue 4 show the slowest mean RTs, as would be expected given the greater number of attribute checks that these pairs require.

Exemplar Model

To test the exemplar model, we fitted Nosofsky and Palmeri's (1997) exemplar-based random walk (EBRW) model to the mean

[7] Although configural cue 0–0 is always a loser, subjects are trained to choose the winning bug on each trial. Although further work is needed to investigate the processes that lead to the observed asymmetry, the pattern of results nevertheless seems intuitively sensible. For example, it is similar in form to the type of visual search asymmetry observed in the classic studies of Treisman and Souther (1985), in which presence of a critical feature leads to pop-out effects, whereas absence of the feature does not.

RT data. The EBRW model is closely related to the GCM. It yields the same predictions of classification choice probability as does the GCM but also generates predictions of RT. According to the model, when a test item is presented, it causes exemplars to be retrieved from memory in accord with how similar the exemplars are to the test item. (Similarity is computed using the same equations as in the GCM.) The retrieved exemplars drive a random walk process. In this process, there is a counter with initial value zero, and the observer establishes criteria that determine the amount of evidence needed for making a Category A or B decision. Each time an exemplar from Category A is retrieved, the counter takes a step in the direction of the A criterion, whereas each time an exemplar from Category B is retrieved the counter takes a step in the direction of the B criterion. The RT is determined by the number of steps required for the random walk to reach either the A or B criterion. For the present task, Category A is composed of the pairs of items that were assigned to the winners category during training, whereas Category B is composed of the losing pairs (i.e., see Version 2 of the exemplar model in The Formal Models section). In general, the model predicts fast RTs in cases in which one item in each pair has been frequently trained as a winner and the other item has been frequently trained as a loser. It predicts slow RTs in cases in which the items in each pair have been trained as winners and losers with roughly equal frequency. The precise predictions, however, depend on how similar the items in each test pair are to previously trained winners and losers as well.

In addition to the overall sensitivity parameter (c) and the attention-weight parameters (w_m) in the GCM, the present applications of the EBRW model made use of a decision-criterion parameter ($crit$), a background-noise parameter (B), a mean residual-time parameter (μ), and a scaling-constant parameter k for transforming the number of steps in the random walk into milliseconds (see Nosofsky & Palmeri, 1997, for a detailed description of the free parameters in the model).[8] As is the case for the cc-TTB model, this version of the EBRW model has eight free parameters.

The fit of the EBRW model is illustrated in Figure 4. It is apparent that the EBRW model performs considerably worse than does the cc-TTB model at predicting the RT data. The model accounts for only 72.2% of the variance in the data (compared with 92.1% of the variance for the cc-TTB model). This best-fitting version of the EBRW model also exhibits various qualitative shortcomings. For example, it predicts virtually identical mean RTs for 11/00 pairs, 01/00 pairs, 10/01 pairs, and 10/11 pairs, but these different pair types show a good deal of variability in the observed data. Likewise, it predicts incorrectly that Cue 3 and Cue 4 pairs should have virtually identical mean RTs.

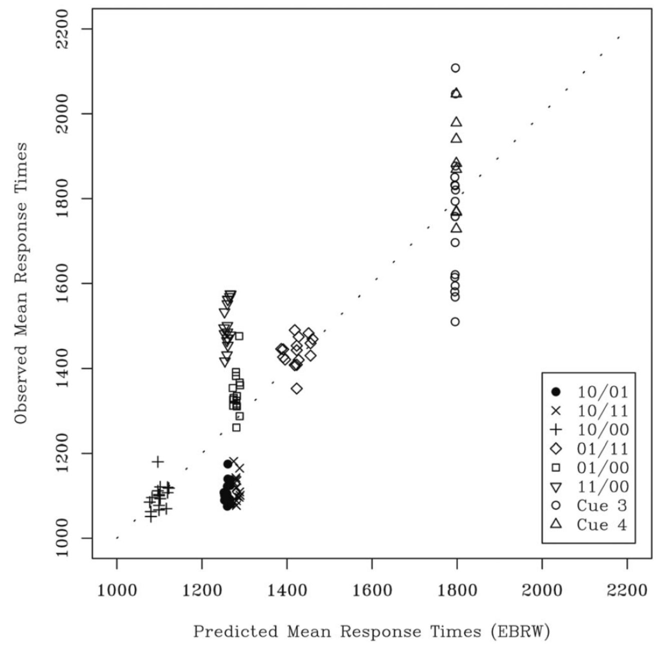

Figure 4. Experiment 3. Mean observed response times for the 120 stimulus pairs plotted against the predictions from the exemplar-based random walk (EBRW) model.

Individual Subject Results

The mean correct RTs for the different stimulus types are reported for the individual subjects in Table 9. (The table does not include the results for Subject 7, who did not continue after Session 1.) As is the case for the grouped mean RT data, the individual subject means are computed across all sessions in which an individual subject scored 95% correct or higher.

To bring out sources of commonality in the individual subject data, we have further summarized the results by computing mean RTs for four main stimulus types. Main Type 1 corresponds to all mismatching pairs involving the 1–0 configural cue (10/01, 10/11, 10/00), whereas Main Type 2 corresponds to all configural-cue mismatches that do not involve cue 1–0 (i.e., 01/11, 01/00, 11/00). Main Types 3 and 4 correspond simply to the Cue 3 and Cue 4 types, respectively. The results for these main types are shown in bold face print in the table.

Inspection of the table reveals that for all 10 subjects, mean RTs for Main Type 1 were faster than for Main Type 2. For 8 of 9 subjects, mean RTs for Main Type 2 were faster than for Main Type 3 (with 1 near tie). And for 8 of 9 subjects, mean RTs for Main Type 3 were faster than for Main Type 4 (with 1 near tie). Pairwise t tests using subjects as the units of analysis confirm that Type 1 RTs were significantly faster than were Type 2 RTs, $t(9) = 5.06$, $p < .001$; Type 2 RTs were significantly faster than were Type 3 RTs, $t(9) = 3.28$, $p < .01$; and Type 3 RTs were significantly faster than were Type 4 RTs, $t(9) = 3.09$, $p < .01$. Thus, the pattern of mean RTs seen in the grouped data is fairly representative of the pattern of RTs observed for most of the individual subjects. Within Main Types 1 and 2, however, there is variability across the individual subjects. For example, some subjects are faster at making 10/01 comparisons, whereas others are faster at making 10/00 comparisons. As we suggested earlier, there is good reason to expect such variability. Multiple competing factors are at work that should affect the time course of comparisons of the configural cues, and it is not surprising to see some individual differences in the relative impact of these factors.

[8] With respect to predicting choice probabilities, the $crit$ parameter in the EBRW model plays the same role as does γ in the GCM. Likewise, the background-noise constant B in the EBRW model plays a similar role as does the guessing parameter g in the GCM. See Nosofsky and Palmeri (1997, pp. 291–292) for more detailed discussion.

Table 9
Mean Response Times for the Stimulus Types for Each Individual Subject in Experiment 3

	Stimulus pair type									
Subject	10/01	10/11	10/00	Main Type 1	01/11	01/00	11/00	Main Type 2	Main Type 3	Main Type 4
1	1176	1065	1094	**1114**	1527	1112	1261	**1297**	**1800**	**2005**
	561	568	568		540	562	562		561	268
2	1475	1470	1637	**1528**	2011	1946	2681	**2208**	**3150**	**3489**
	361	381	376		369	377	363		332	162
3	983	975	923	**960**	1239	963	1196	**1131**	**1435**	**1271**
	431	440	445		429	443	427		386	214
4	1089	966	820	**955**	1144	935	963	**1008**	**1003**	**1224**
	174	185	190		164	189	187		187	91
5	819	916	762	**832**	1675	1259	1408	**1436**	**1415**	**1708**
	187	190	192		163	191	185		176	88
6	1273	1293	1140	**1234**	1279	1480	1615	**1457**	**1768**	**1839**
	493	493	508		490	500	480		486	245
8	1091	1203	1036	**1110**	1336	1591	1643	**1522**	**1969**	**2130**
	373	376	381		375	368	368		371	186
9	1074	1051	1181	**1102**	1233	1308	1277	**1273**	**1354**	**1505**
	570	572	565		561	549	553		536	271
10	745	950	804	**834**	1278	1143	1225	**1215**	**1327**	**1490**
	305	312	308		290	294	304		287	147
11	1097	1103	1219	**1140**	1823	1571	1700	**1696**	**2025**	**2018**
	318	316	317		303	316	316		278	150

Note. Top line in each row = mean response time; bottom line in each row = number of observations. Main Type 1 = stimulus pairs involving a mismatch on configural cue 1-0. Main Type 2 = configural-cue mismatches not involving configural cue 1-0. Main Type 3 = stimulus pairs involving a match on the configural cue and a mismatch on Cue 3. Main Type 4 = stimulus pairs involving a match on the configural cue, a match on Cue 3, and a mismatch on Cue 4. The results for the main types are in bold print.

We also fitted the cc-TTB and EBRW models to the individual subject mean RTs. The results are reported in Table 10. The cc-TTB model provides better fits than does the EBRW model for all 10 subjects. The log-transformed root mean squared deviations yielded by the cc-TTB model are significantly smaller than the ones yielded by the EBRW model, $t(9) = 6.70$, $p < .0001$.

cc-TTB Fits to the Experiment 2 Choice-Probability Data

In light of these RT results favoring the cc-TTB model, we decided to fit a version of the cc-TTB model to the choice-probability data of the top performers from our Experiment 2. The model was the same as the gTTB model already described, except the separate independent-cue weights for Attributes 1 and 2 were replaced by a single configural-cue weight for these attributes. Furthermore, for simplicity, we assumed that the top performers had learned the complete ordering defined by these attribute combinations (i.e., that configural cue 1–0 predicted the highest values on the criterion variable, 0–1 the next highest values, and so forth). The summary BIC fits for the top performers are shown along with those of the gTTB and exemplar models in Table 7. For 19 of the 21 top performers, the cc-TTB model provides a better quantitative fit to the data than does the exemplar model. Thus, beyond fitting RTs, the cc-TTB model can also provide good accounts of choice-probability data in situations in which subjects develop sensitivity to cue interactions.

Table 10
Fits of the Models to the Individual Subject Data in Experiment 3

	Model			
	cc-TTB		EBRW	
Subject	RMSD	%Var	RMSD	%Var
1	64.6	95.7	83.6	92.7
2	299.6	83.6	473.2	59.2
3	137.7	63.3	157.6	51.9
4	94.7	58.2	101.7	51.8
5	217.4	72.1	255.7	61.4
6	122.1	77.6	170.8	56.2
8	197.8	76.2	272.7	54.8
9	68.5	77.5	109.6	42.4
10	146.7	72.2	199.0	48.7
11	200.3	77.5	256.2	63.1

Note. cc-TTB = configural-cue take-the-best model; EBRW = exemplar-based random walk model; RMSD = root mean squared deviation between predicted and observed mean response times; %Var = percentage of variance accounted for.

Summary

In sum, our modeling analyses of the RT data from Experiment 3 and revised modeling of the choice-probability data of the top performers from Experiment 2 yield results that favor a configural-cue version of the TTB model and that again challenge the predictions from exemplar-based models of multi-attribute paired-comparison inference.

General Discussion

This research was motivated by the idea of testing exemplar-based models of categorization in the domain of multi-attribute paired-comparison inference. Although there has been some past consideration of exemplar models in this domain (e.g., Juslin & Persson, 2002; Juslin, Jones, et al., 2003), the focus was on situations in which observers were provided with continuous-valued feedback of the criterion variable and needed to learn continuous-valued functional relations to perform the tasks. By contrast, in the present research, observers were provided with only discrete trial-by-trial feedback as to which member of training pairs had a higher value on the criterion variable. This type of discrete, categorical feedback moves the paradigm much closer to the type of categorization learning for which exemplar models were originally formalized (e.g., Medin & Schaffer, 1978). Furthermore, given the well-known success of exemplar models in the domain of categorization, it is only natural to test them in the present closely related domain.

It seems intuitively reasonable that exemplar-based storage and retrieval processes might play a major role in this type of learning environment. During the course of learning, participants received feedback that certain exemplars tend to be high on the criterion variable, whereas others tend to be low. Thus, a straightforward mechanism for making multi-attribute inference judgments would be to assess the similarity of members of test pairs to the experienced exemplars. Items that are similar to exemplars that have been trained as being high on the criterion variable would tend to be judged as high on the criterion as well. Indeed, we found strong initial support for the exemplar-based modeling ideas in some preliminary model-fitting tests. In particular, in a domain adapted from a natural, real-world environment, we found that formalized versions of the exemplar model yielded quantitative fits to data that were as good or better than those of alternative classic models of multi-attribute paired-comparison inference, namely the gTTB and WADD models of judgment.

Our key idea for contrasting the predictions from the exemplar model with those of the alternative models was to develop structures that required sensitivity to correlated attributes in the training pairs. As explained earlier, the standard versions of the TTB and WADD models can be viewed as independent-cue models, in which information from individual attributes contributes in an independent manner to the overall judgment of the criterion variable. By contrast, the exemplar model is an interactive-cue model and predicts sensitivity to correlated-cue structure. Thus, we developed test structures in which individual cues provided little independent information about the magnitude of the criterion variable, whereas cue configurations provided a great deal of information. Therefore, independent-cue models such as TTB and WADD predicted exceedingly poor performance in such tasks, whereas the interactive-cue exemplar model predicted good performance.

To develop rigorous tests of these predictions, in Experiments 1 and 2 we used a method based on generalization testing (Busemeyer & Wang, 2000). First, we fitted the alternative models to the paired-comparison inference data obtained in the standard, independent-cue experiment of Bergert and Nosofsky (2007). (This previous experiment used the same types of stimuli, number of training trials, training and testing procedures, and population of subjects as did the current experiments.) Next, we held fixed the levels of certain overall-performance parameters (sensitivity, response-scaling, and guessing) that were estimated from those fits. Finally, we combined those parameter estimates with a variety of plausible attribute-weighting schemes that observers might apply for the present correlated-cue tasks. Regardless of the details of the attribute-weighting scheme, the exemplar models predicted distributions of performance in a different ballpark (i.e., far better) than what was observed for the vast majority of participants in our studies. This same pattern of misprediction held for both Experiment 1, which involved correlated cues defined over two separate pairs of attributes, and Experiment 2, which involved correlated cues defined over just a single pair of attributes. By contrast, in both experiments, the vast majority of subjects showed performance at levels predicted by the independent-cue strategies.

Furthermore, we obtained converging evidence for this pattern of results by demonstrating that, even with all parameters allowed to vary freely, the exemplar models yielded worse quantitative fits than did the independent-cue models to the data of the vast majority of individual subjects.

Nevertheless, even in the present Experiments 1 and 2, a small subset of subjects exceeded the performance limits allowed by use of an independent-cue TTB or WADD strategy, suggesting that such subjects developed some sensitivity to the correlated-cue structure. One possibility is that, after learning that independent-cue strategies such as TTB or WADD were inadequate, such subjects were making a transition to an alternative strategy (cf. Johansen & Palmeri, 2002; Juslin, Olsson, & Olsson, 2003). To investigate this idea, in Experiment 3 we tested a smaller group of subjects on the same structure used in Experiment 2, except now subjects participated for 8–10 sessions of training, instead of just a single session. Furthermore, we recorded subjects' RTs for making their paired-comparison decisions to gain information about the strategy that they might be using.

Under these extended training conditions, the subjects' performance skyrocketed, and most came to achieve nearly perfect accuracy in the task. However, detailed modeling analyses of the RT data suggested that the subjects did not switch to the hypothesized exemplar-based coding strategy for making their paired-comparison decisions. Instead, a more viable explanation of the pattern of RT results is that subjects learned to recode the correlated attributes into a higher order configural cue. They then engaged in a series of rule-based tests involving the recoded cues, much in the spirit of a TTB process (cf. Garcia-Retamero et al., 2007). Specifically, the mean RT data were well predicted by a model that assumed that subjects first compared the pairs of objects along the highly diagnostic configural cue. If there was a mismatch, then an immediate decision could be made. If there was a match, then the next most diagnostic attribute determined the direction of choice and so forth until a mismatching attribute was found. By contrast, the exemplar model provided much worse fits to the mean RT data. This superiority of the cc-TTB model held at both the level of grouped mean RTs as well as at the level of individual subjects.

Although the good fits of the cc-TTB model provide support for the general idea of a decision-making strategy based on a set of hierarchically organized rules (some involving configural cues), future research is needed to flesh out more details of the process. For example, although the pattern of best-fitting parameter estimates from this model seemed intuitively sensible, we need to investigate the processes that determine the speed with which alternative configural cues can be compared as well as how the learned strengths of the configural cues map onto RTs. Likewise, the processes that determine the speed with which matching and mismatching attributes can be located in these kinds of perceptual displays also need to be better understood.

Given the well-known success of exemplar models in the domain of categorization and the frequent demonstrations of sensitivity to correlated-cue structure in such tasks (Elio & Anderson, 1981; Juslin, Jones, et al., 2003; McKinley & Nosofsky, 1995; Medin, Altom, Edelson, & Freko, 1982; Medin & Schaffer, 1978; Nosofsky, 1987; Nosofsky, 1992), we found the present results to be extremely surprising. Is there indeed a sharp discontinuity between the processes of multi-attribute categorization and multi-attribute prediction of ordered criterion variables? Certainly, the present paired-comparison inference tasks required observers to establish far more fine-grained distinctions in memory than do typical tasks of categorization. For example, whereas typical tasks of categorization require observers to classify objects into only a few nominally labeled groups, the present tasks required observers to learn the precise ordering of 16 items. Furthermore, the number of training pairs in the inference tasks is far greater as well. For example, in a structure with 16 stimuli, there are 16 training items in a task of categorization but 120 training pairs in the present tasks of paired-comparison inference. These vastly different computational demands are almost certainly a key part of the story.

From another perspective, however, there may be more continuity between the tasks than is readily apparent. For example, even in categorization, although the end result is often one that shows evidence of exemplar-based storage and retrieval, research suggests that at early stages of learning observers rely on use of single-attribute rules (e.g., Johansen & Palmeri, 2002; Nosofsky, Palmeri, & McKinley, 1994; Nosofsky & Zaki, 2002). Thus, it is interesting to hypothesize that the same progression of processing strategies may operate in both tasks. Because of the reduced computational demands, the progression can take place more quickly in typical tasks of categorization than in the present tasks of multi-attribute paired-comparison inference. Although the best current account of our extended training results from Experiment 3 involves the development and application of configural-cue rules, it should be noted that application of such a strategy allowed for perfect and highly efficient performance in the particular paradigm that was tested. With extended training on more complex structures, perhaps a role of stored exemplars in multi-attribute paired-comparison decision making may yet be observed.

References

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

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 107–129.

Bröder, A., & Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132,* 277–293.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44,* 171–189.

Cohen, A. L., & Nosofsky, R. M. (2000). An exemplar-retrieval model of speeded same–different judgments. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1549–1569.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, and the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York: Oxford University Press.

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.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 397–418.

Erickson, M. A. (1999). Rules and exemplars in category learning (Doctoral dissertation, Indiana University, Bloomington, 1999). *Dissertation Abstracts International, 60,* 2377B.

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

Garcia-Retamero, R., Hoffrage, U., Dieckmann, A., & Ramos, M. (2007). Compound cue processing within the fast and frugal heuristics approach in nonlinearly separable environments. *Learning and Motivation, 38,* 16–34.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox.* Cambridge, MA: MIT Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, and the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Heit, E., Price, P., & Bower, G. (1994). A model for predicting the outcomes of basketball games. *Applied Cognitive Psychology, 8,* 621–639.

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

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482–553.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 924–941.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in multiple-cue judgment. *Journal of Experimental Psychology: General, 132,* 133–156.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review, 11,* 343–352.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12,* 605–621.

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

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences, USA, 97,* 11170–11175.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences, 9,* 11–15.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 53–65.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–109.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M., Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

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

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

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

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 438–469). Cambridge, England: Cambridge University Press.

Schwarz, G. (1978). Estimating the attribute of a model. *Annals of Statistics, 6,* 461–464.

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

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114,* 285–310.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79,* 281–299.

Appendix A

Modified Version of the Generalized Take-the-Best Model for Application in the Correlated-Cue Paradigms

In the modified version of the generalized take-the best (gTTB) model, the probability that Alternative A is chosen from pair (A,B) is given by

$$P(A;A,B) = \frac{\sum_{a \in FA} w_a + g}{\sum_{a \in FA} w_a + \sum_{b \in FB} w_b + 2g}, \qquad \text{(A1)}$$

where the parameters w_a and w_b are as defined in the text, FA denotes the set of attributes that discriminate in favor of Alternative A (and likewise for FB), and where g ($0 \leq g$) is a guessing-weight parameter. Note that if the value of g is large relative to some of the attribute weights, then observers tend to guess even if the values associated with those attributes discriminated between the alternatives. Finally, as noted in the text, the modified model was allowed to choose the direction of choice for each attribute value. By contrast, in previous applications of gTTB, the cue value that tended to point to larger values of the criterion variable was always defined as the positive cue value. This assumption is not workable in the present context because in many cases the cue validities of the individual attribute values are .50 or very close to .50.

Appendix B

Mean Response Times for All Pairs of Stimuli Used in Experiment 3

Poison	Stimulus	0	1	10	11	15	16	25	26	60	61	70	71	75	76	85	86
		0000	0001	0010	0011	1100	1101	1110	1111	0100	0101	0110	0111	1000	1001	1010	1011
0	0000		2047	1758	1851	1486	1432	1482	1496	1260	1392	1312	1354	1103	1112	1062	1085
1	0001	2047		2047	1697	1454	1501	1417	1533	1335	1313	1331	1327	1093	1100	1096	1051
10	0010	1758	2047		1940	1479	1569	1553	1465	1287	1476	1324	1382	1120	1069	1077	1180
11	0011	1851	1697	1940		1574	1575	1563	1471	1361	1367	1311	1317	1118	1107	1121	1066
15	1100	1486	1454	1479	1574		1884	1831	1832	1353	1490	1445	1446	1077	1138	1132	1086
16	1101	1432	1501	1569	1575	1884		2108	1794	1420	1453	1421	1426	1099	1118	1181	1087
25	1110	1482	1417	1553	1563	1831	2108		1978	1459	1483	1409	1408	1109	1096	1120	1143
26	1111	1496	1533	1465	1471	1832	1794	1978		1469	1430	1475	1443	1101	1165	1136	1088
60	0100	1260	1335	1287	1361	1353	1420	1459	1469		1768	1569	1580	1075	1105	1101	1108
61	0101	1392	1313	1476	1367	1490	1453	1483	1430	1768		1877	1621	1175	1089	1098	1089
70	0110	1312	1331	1324	1311	1445	1421	1409	1475	1569	1877		1770	1124	1091	1122	1099
71	0111	1354	1327	1382	1317	1446	1426	1408	1443	1580	1621	1770		1080	1138	1140	1099
75	1000	1103	1093	1120	1118	1077	1099	1109	1101	1075	1175	1124	1080		1869	1594	1614
76	1001	1112	1100	1069	1107	1138	1118	1096	1165	1105	1089	1091	1138	1869		1820	1510
85	1010	1062	1096	1077	1121	1132	1181	1120	1136	1101	1098	1122	1140	1594	1820		1729
86	1011	1085	1051	1180	1066	1086	1087	1143	1088	1108	1089	1099	1099	1614	1510	1729	

Note. Stimulus codings do not include Attributes 5 and 6, which are constant across all stimuli.

Received June 28, 2006
Revision received June 15, 2007
Accepted July 30, 2007 ■