Psychonomic Bulletin & Review
2001, 8 (4), 629-647

A hybrid model of categorization

JOHN R. ANDERSON and JONATHAN BETZ
Carnegie Mellon University, Pittsburgh, Pennsylvania

Category learning is often modeled as either an exemplar-based or a rule-based process. This paper shows that both strategies can be combined in a cognitive architecture that was developed to model other task domains. Variations on the exemplar-based random walk (EBRW) model of Nosofsky and Palmeri (1997b) and the rule-plus-exception (RULEX) rule-based model of Nosofsky, Palmeri, and McKinley (1994) were implemented in the ACT-R cognitive architecture. The architecture allows the two strategies to be mixed to produce classification behavior. The combined system reproduces latency, learning, and generalization data from three category-learning experiments—Nosofsky and Palmeri (1997b), Nosofsky et al., and Erickson and Kruschke (1998). It is concluded that EBRW and ACT-R have different but equivalent means of incorporating similarity and practice. In addition, ACT-R brings a theory of strategy selection that enables the exemplar and the rule-based strategies to be mixed.

Research on human category learning has a history that extends back at least to Hull's (1920) study of learning to categorize Chinese symbols and his conclusions in favor of an associative learning proposal. It was an important domain early in the cognitive revolution, when theorists argued for various hypothesis-testing theories (e.g., Bower & Trabasso, 1963; Levine, 1975). The hypothesis-testing theories were based on research with stimuli that had a very simple, often one-dimensional categorical structure. The 1970s saw a renewed interest in more complex, fuzzy categories and in proposals for prototype theories (Reed, 1972; Rosch, 1975) and exemplar theories (e.g., Medin & Schaffer, 1978). The rise of connectionist models resulted in the proposal of associative theories (e.g., Gluck & Bower, 1988) not that different from the original Hull proposal. Whereas the original research focused on accuracy data, there has been a new emphasis on latency data, to help choose among theories (e.g., Lamberts, 1998; Nosofsky & Palmeri, 1997a). Recently, neuroimaging and other cognitive neuroscience data have been recruited in order to try to decide among alternative theories (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; E. E. Smith, Patalano, & Jonides, 1998). There has been an impressive growth in the characterizations of the phenomena in category learning. However, the field does not seem to be coming any closer to a consensus as to what *the* mechanism of category learning is.

This paper is based on the assumption that this contest between alternative theories has been cast too narrowly, in two different senses. First, this contest has been too narrow in that categorization learning may not be the outcome of a single mechanism. There is a new, emerging view that categorization behavior might be some mix of different methods of categorization (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; E. E. Smith et al., 1998; J. D. Smith, Murray, & Minda, 1997). Erickson and Kruschke (1998) proposed that the outputs of an exemplar and a rule module are mixed to produce a final result. Ashby et al. and J. D. Smith et al. proposed an alternative possibility, which is that, on a trial-by-trial basis, participants choose to use either an implicit system or a verbal rule system. Both the Erickson and Kruschke (1998) and the Ashby et al. models are basically connectionist systems.

The second way this contest has been narrow is that it has tended to focus only on categorization data. There is little reason to believe that categorization learning is an isolated cognitive process. The same mechanisms that are involved in categorization should be involved in other cognitive processes, and the same mechanisms that are responsible for category learning should be involved in learning other knowledge. Most researchers in the field seem to accept this, at least implicitly. For instance, Nosofsky and Palmeri (1997b) justify aspects of their model with results on perceptual identification. The cognitive neuroscience literature that Ashby et al. (1998) cite often involves noncategorization tasks. Although there may be informal agreement on this, up to now there has been no effort to formally integrate categorization into a general set of mechanisms that apply in mul200tiple domains. Constraining models of categorization to be more generally consistent with cognition eliminates many degrees of freedom in the formulation of the theories. For instance, the processes of memory failure and the timing of cognitive steps should be the same in a model of cognition as in models of other cognitive tasks. Moreover, if we are going to mix different methods of categorization in a hybrid model and we do not want our degrees of freedom to multiply, the parameters that

This research was supported by Grant N00014-96-1-0491 from the Office of Naval Research. We thank Christian Lebiere, Robert Nosofsky, Thomas Palmeri, Alex Petrov, and Lynne Reder for their comments on this paper. Correspondence concerning this article should be addressed to J. R. Anderson, Department of Psychology, BH345D, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (e-mail: ja@cmu.edu).

govern one method should be the same as the parameters that govern another method.

This paper will present a model of categorization that implements two strategies for categorization in the same system, ACT-R (Anderson & Lebiere, 1998). ACT-R is a cognitive architecture whose basic mechanisms and processes have been determined by the development of models in such domains as verbal learning, strategy choice, cognitive arithmetic, analogy, and scientific reasoning. However, until this paper, it has not been applied to the domain of categorization learning. Besides constraining the model of categorization to operate in a way consistent with models of other phenomena, ACT-R provides a theory of how participants choose among multiple strategies for categorization.

The model in this paper will combine a rule-based submodel and an exemplar submodel. Our submodels will be based on two relatively successful models proposed by the same authors—Nosofsky, Palmeri, and McKinley's (1994) rule-plus-exception (RULEX) model and Nosofsky and Palmeri's (1997b) exemplar-based random walk (EBRW) model. There were two motivations for trying to use existing models, rather than creating models from scratch. One is that it reduces the possibility that the success of this effort could be produced by clever tricks we put into our made-up models. The other is that it will avoid unnecessarily complicating the field with more models that differ from each other in ways whose importance is unclear. Rather, by showing how these models can be implemented in ACT-R, we will contribute to convergence in the field by showing how the mechanisms in these models relate to the mechanisms in ACT-R.

It is not by any means a foregone conclusion that we will be able to implement either of these models in ACT-R, and we regard it as a contribution to show that we can. As we will argue in the conclusion, many architectures, including past versions of ACT (e.g., Anderson, 1983), would not be capable of this. Moreover, there are models of categorization that could not be implemented in ACT-R—particularly, many of the connectionist models. To the extent that RULEX and EBRW are successful models, implementing them in ACT-R extends credit to ACT-R by showing that it is compatible with the data on categorization. To the extent ACT-R has provided successful models of other domains, it extends credit to these models by showing that they are generally compatible with what is known about cognition.

This effort makes a second contribution that goes beyond showing a general compatibility between ACT-R and these categorization models. This is that ACT-R has an explanation for how people choose between two bases, such as rules and exemplars, for categorizing stimuli. According to the ACT-R theory of choice (Lovett, 1998), participants track how well each basis is working on the stimulus set and select each method in rough proportion to its past history of success (see also Reder, 1987, 1988). This is a learning mechanism that ACT-R brings to the table that has not been part of existing theories of categorization.

It is worth noting from the outset that this paper is sidestepping the traditional contest of models. We will not be claiming that there is some new datum that uniquely leads us to prefer the hybrid ACT-R model over other models in the arena. We will be content to show that it does as well as other models. The search for the decisive data set has brought a lot of enlightenment to category learning, but it has not succeeded in identifying the correct model. We think that it is an equally important contribution to show that a model is more generally compatible with what is known about human cognition.

The Exemplar-Based Random Walk Model

The EBRW model combines major properties of Nosofsky's (1988) generalized context model of classification (GCM) and Logan's (1988) instance-based model of automaticity. According to the GCM, participants store the individual exemplars of the categories as they study them. Participants tend to classify a stimulus into the category of the instances to which it is most similar. Logan's model describes how people learn to perform skilled actions. Performance of skilled actions initially depends on using some algorithm. Each time the action is successfully completed, an instance is stored in memory. Later, these stored instances can be recalled and used to perform the task. Skilled performance, then, is a race between executing the algorithm for the task and recalling prior instances. Experience with a skilled action leads to storage of many instances, and eventually these instances are used more than the initial algorithm. This model accounts for the power-law decreases in reaction time observed with training.[1]

The EBRW model combines the GCM concept of category learning as comparison with stored exemplars with Logan's (1988) concept of a race among stored instances. Unlike Logan's model, all stored instances race to be retrieved, instead of just those that are identical to the presented stimulus. The speed at which an exemplar is recalled is proportional to its similarity to the presented stimulus. The exemplar with the fastest retrieval time is used to assign a category to the stimulus. In this way, the exemplars that have been used more frequently and are more similar to the presented stimulus are more likely to affect its categorization decision.

After an exemplar is retrieved, an internal counter is updated on the basis of the category of the retrieved exemplar. For example, consider the situation of choosing between two categories. The internal counter would begin at 0. Retrieved exemplars from one category would cause positive increments, whereas exemplars from the other category would cause negative increments. When the counter exceeds a threshold absolute value, the model categorizes the stimulus into either the positive or the negative category.

The two factors that determine classification time are the number of steps in the random walk and the speed of each step. The EBRW model predicts that stimuli that are very similar to instances in one category and very dissimilar to instances in other categories should have rapid clas-

sification times. Such stimuli would most often provoke recalls from exemplars in only one category, so the random walk would move consistently in one direction. Also, stimuli that are similar to stored exemplars from different categories should show slower response times, because such stimuli would provoke recalls of exemplars in different categories, and therefore, the random walk would vacillate. The EBRW model also predicts that increased experience with stimuli will decrease reaction time. As more exemplars are stored in memory, the retrieval time that wins the race in each step of the random walk will be faster, so the total time for the random walk to complete will also be faster.

The Rule-Plus-Exception Model

The RULEX model paints a very different picture of categorization. This model searches through the space of possible rules to classify stimuli. Rules are tested one by one, until a rule is found that meets a performance criterion. Exceptions to rules can be stored to account for stimuli that are incorrectly classified by the chosen rule.

Search begins with rules that classify stimuli according to a single dimension—for example, that an item is in a category if it is red. Initially, RULEX tries to find perfect rules that can classify all stimuli without exceptions. A perfect rule is discarded when feedback indicates that it has produced an incorrect category judgment. If a perfect rule lasts through an entire block of trials, it is kept permanently, and there is no need for further search. If all perfect rules are eliminated, search continues with single-dimension, imperfect rules that are not required to classify all the instances. Each single-dimension imperfect rule is tested for a number of trials, usually one block. If the rule satisfies a lax criterion of accuracy over this test window, it is tested according to a stricter criterion for some number of trials, usually another block of trials. If the rule passes this stricter criterion of accuracy, it is kept permanently. At this point, the system begins learning exceptions, to counter the mistakes of the imperfect rule. If the rule does not pass the stricter criterion, it is discarded, and another rule is selected. If no single-dimension imperfect rule satisfies the stricter criterion, search continues with imperfect, conjunctive rules. These rules are tested in a manner similar to single-dimension imperfect rules. If no rule passes the stricter criterion, only a set of exceptions are stored. However, it never gets this far in the experiments that we will consider.

Exceptions are formed when a permanent rule makes a misclassification of a stimulus. A stored exception is an association between an incomplete pattern and a category label. For example, with four-dimension binary stimuli, if the permanent rule is "The value 2 on Dimension 3 indicates Category B" (denoted **2* $\Rightarrow$ B), a possible exception would be *12* $\Rightarrow$ A. In forming exceptions, any dimensions used in the permanent rule are used, and all the remaining dimensions are sampled with a fixed probability. If an exception leads to making an incorrect category decision, it is eliminated.

When the model needs to make a category decision about a presented stimulus, it first checks for any applicable exceptions. If there are multiple exceptions that apply, the exception that specifies values for the most dimensions is used. If no exceptions apply, a judgment is made according to the current rule.

A primary prediction of the RULEX model concerns the transfer pattern of responses made to novel stimuli. According to RULEX, through random choices, different participants will come to different rules that lead to different transfer patterns. However, some rules are much more likely, given the way RULEX searches its rule space. The transfer patterns that are due to these rules will occur most frequently.

ACT-R

Before describing the ACT-R model for categorization, we need to describe some features of the ACT-R architecture. According to ACT-R, cognition emerges from the interaction of two types of knowledge—declarative knowledge that encodes explicit facts that the system knows and procedural knowledge that encodes rules for processing declarative knowledge. In ACT-R, information processing is under the control of a current goal. In response to that goal, a production rule is chosen from procedural memory for application. Typically, a production rule will call for the retrieval of some piece of information, called a chunk, from declarative memory, which will result in a transformation of the goal. Then, the cycle of production selection and information retrieval will apply to this new goal state. Two aspects of ACT-R that are important for present purposes are the process by which production rules are selected to apply to the goal and the process by which chunks are selected to be retrieved.

Selection of Production Rules

Conflict resolution is the term used to refer to selection among production rules. A good illustration of conflict resolution occurs in Lebiere's (1998) model of the development of arithmetic knowledge, which bears a good number of similarities to Logan's (1988) model of skill acquisition. In Lebiere's model (as in many models of cognitive arithmetic—e.g., Ashcraft, 1995; Campbell, 1997; Reder & Ritter, 1992; Siegler, 1988), when a child is faced with the goal of adding two numbers, such as 4 and 3, there are two strategies that can apply. In Lebiere's model, these two strategies are implemented as production rules. One rule calls for a retrieval of the information:

IF the goal is to find a + b
  and c can be retrieved as the sum of a and b
THEN the answer is c.

The other rule sets a subgoal to try back-up computation:

IF the goal is to find a + b
THEN set a subgoal to count b units past a,

after which a series of productions will compute the answer. The first production rule has the advantage that it can produce the answer faster, because it does not call on a counting subprocedure. However, the model may fail to retrieve anything, and it will have to go on to the back-up computation. Also, the model may retrieve the wrong answer. ACT-R's selection among such production rules is determined by their expected utility, which ACT-R calculates as $PG - C$, where P is the expected probability that the goal will succeed, G is the value of the goal, and C is the expected cost of the that rule. In this paper, we will use the ACT-R defaults of 20 for G and of measuring C as the time in seconds for the goal to be achieved.

These utilities are noisy, real-valued quantities in ACT-R. Because of noise, a normally lower valued utility will sometimes become larger than a normally higher valued utility. The production chosen on a particular trial will be the production that, among those matching the goal, has the highest utility on that trial. The probability of that happening will be a function of the mean utility, E_i, of that production, the mean utilities of competing productions, and the noise in utility. The formula describing this is

$$\text{Probability of choosing } i = \frac{e^{E_i/t_E}}{\sum_j e^{E_j/t_E}},$$

(Conflict Resolution, Equation 1)

where the summation in the denominator is over the various productions j that might apply. This is a "soft max" rule, which tends to select the production with maximum utility, but not always because the utilities are noisy and can reverse on a particular trial. The parameter t_E in the above distribution is related to the standard deviation, σ_E, of the noise by the formula $t_E = \sqrt{6}\sigma_E/\pi$. This equation is the same as the Boltzmann equation used in Boltzmann machines (Ackley, Hinton, & Sejnowsky, 1985; Hinton & Sejnowsky, 1986). In this Botlzmann machine context, t_E is called the temperature. Throughout the models in the paper, our estimate of t_E will be 2.2.

This process of selecting productions will be key in our ACT-R model for categorization. Production rules will embody three competing ways of classifying a stimulus. One is by implementing the EBRW strategy. The second is by implementing the RULEX strategy. The third, which will be the only one applicable at the beginning, is by guessing.[2] On any trial, ACT-R will choose to pursue one of these strategies. With experience, one of these strategies will tend to have the most success and, therefore, will be chosen most often. The characteristics of various domains may force ACT-R to configure itself to behave exclusively according to one of the strategies. The system is designed to have a high utility for using rules initially. This is because it takes some time to discover successful rules, and so it needs to have this bias to make it persevere until a successful rule can be found. This setting of the model, to prefer rules, can be interpreted as reflecting a bias that categories are rule-based, and this corresponds to a naive classical view of category structure (E. E. Smith, 1989). The bias means ACT-R will transition from guessing to rule-based classification if it can form rules. If it cannot form rules or if they prove unsuccessful, it will have to resort to exemplar-based classification. As we will also see, sometimes even though it finds an adequate set of rules, it may switch to exemplar-based classification after extensive practice, because this proves faster.

Retrieval of Declarative Chunks

Declarative knowledge in ACT-R is represented in units called *chunks*. In our model, the relevant chunks are ones that encode exemplars and categorization rules. Thus, ACT-R can have chunks encoding both the example "the large red triangle was in category A" and the rule "red objects are in category A." Retrieval of declarative chunks from memory is probabilistic, like conflict resolution. This probabilistic retrieval can be illustrated in the arithmetic domain, where chunks will encode facts like 3 + 4 = 7. When faced with the problem 3 + 4, the child will tend to retrieve the correct 3 + 4 = 7 chunk, but other chunks might be retrieved. The child may fail to retrieve anything or may retrieve similar chunks, like 3 + 5 = 8 or 3 ∗ 4 = 12, and produce the wrong answer. Or, as Siegler (1988) has argued, the child may have once solved 3 + 4 with the answer 6, and now the 3 + 4 = 6 fact is a weak chunk in the database, which can intrude.

According to ACT-R, the selection among different chunks is determined by their levels of activation. For our purposes, two factors are relevant in determining a chunk's level of activation. One is the amount of past practice that the chunk has had. ACT-R assumes that activation increases as a logarithmic function of amount of practice. Second, activation will reflect the degree of match between the chunk and the retrieval specifications. For example, the chunk encoding 3 ∗ 4 = 12 will mismatch a 3 + 4 retrieval probe with respect to the plus operator and so will not be as active as the 3 + 4 = 7 chunk. In general, ACT-R calculates a mismatch by summing the differences between all the elements (3, +, and 4, in the example) in the retrieval probe and the values in the declarative chunk. This partial matching process will become particularly important in applying ACT-R to categorization tasks with continuously varying dimensions, since it is possible that no memory chunk exactly matches the stimulus to be categorized. The formula giving the activation, A_i, of chunk i is

$$A_i = \ln N_i - M_i, \quad \text{(Activation, Equation 2)}$$

where N_i is the number of trials of practice and M_i is the degree of mismatch of chunk i to the production.[3] Note that the effect of practice is implemented differently in ACT-R than in EBRW or Logan's (1988) model. In the EBRW model and Logan's model, each repetition of an example causes another instance to be stored. In ACT-R, when an example is repeated, the base-level activation for the sin-

gle chunk encoding the example is higher. Nonetheless, as others have noted (e.g., Nosofsky & Alfonso-Reese, 1999; Wixted, Ghadisha, & Vera, 1997), models that assume strengthening of a single trace can be equivalent in their predictions to models that assume that repetitions create new encodings.

Just as with utilities, activations are noisy, real-valued quantities. ACT-R will retrieve the most active chunk if it is above a minimum threshold of activation, but this can vary from retrieval to retrieval. The actual predictions of the model are obtained by Monte Carlo simulations, but two equations have been shown (Anderson & Lebiere, 1998) to give good approximate characterizations of the model's behavior. The probability of retrieving a chunk i is given by the formula

$$\text{Probability of retrieving } i = \frac{e^{A_i/t_A}}{\sum_j e^{A_i/t_A}},$$

(Chunk Choice, Equation 3)

where the summation is over all possible chunks j. This is basically the same soft-max equation as in production choice, and t_A reflects the noise in the activation values.[4] ACT-R will retrieve a chunk only if it is above an activation threshold. The probability that a chunk will fall above a threshold of activation is given by

$$\text{Probability} = \frac{1}{1+e^{-(A_i-\tau)/s_A}},$$

(Retrieval Probability, Equation 4)

where τ is the threshold and s_A is also related to the standard deviation, σ_A, of the activation noise and to t_A by the formula $s_A = \sqrt{3}\sigma/\pi = t_A/\sqrt{2}$. The value of the activation noise s_A was kept constant in these simulations at .55 (and so t_A was constant at .78). This is comparable with values used in other simulations (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998). Finally, the time to retrieve a chunk is described by ACT-R's retrieval time equation:

$$\text{Time} = Fe^{-A_i}. \quad \text{(Retrieval Time, Equation 5)}$$

In the simulations, we set the latency scale factor F to 1.0 sec, which is a typical (and default) value.

Within the general ACT-R architecture just sketched, a system was created that implemented both the EBRW model and the RULEX model. Basically, this involved having production rules direct the selection and execution of categorization strategies whereas declarative information encoded the examples and rules that provided the category information.

THE ACT-R HYBRID MODEL

As we noted earlier, ACT-R has three ways of classifying a stimulus: guessing, using a rule module based on RULEX, and using an exemplar module based on EBRW. The actual simulation resides at the *Published Models* link at the ACT-R home page (http://act.psy.cmu.edu/), where one can inspect the code and try various parameter combinations for the data sets described later. However, here we will try to explain the basic principles by which these three modules operate. Table 1 gives snippets of production rule firings that illustrate each of the methods classifying a stimulus. Each of these snippets presents the cycle number denoting the sequence of that production firing in a larger run and the time in seconds of that production firing. Although no method is always correct, for comparability the snippets chosen show the methods correctly classifying a stimulus. The first snippet, illustrating guessing, is the simplest. The first production issues a guess, and the second processes the feedback that the guess was correct. Then follows a sequence of three rules that fire after every successful classification event. The first encodes the association of the four features of the stimulus with the category. This will provide an exemplar for later use by the exemplar module. The remaining two terminate this study and note that the overall trial was successful. The choice of guessing in this example occurred only because there were no rules or exemplars available at the early point in the run from which this snippet comes. The guess option is rated so low that it is never chosen when an exemplar or a rule can apply and it never achieves enough success to make it preferable. In the subsections to follow, we will describe in more detail the implementation of the rule and exemplar modules in ACT-R.

Implementing EBRW in ACT-R

The EBRW subsystem of the ACT-R hybrid model differed slightly from the original EBRW model, owing to differences in similarity evaluation and the structure of declarative memory. In Nosofsky and Palmeri's (1997b) original formulation of the EBRW model, the similarity between two exemplars i and j is calculated as the weighted Euclidean distance between the two:

$$d_{ij} = \sqrt{\sum_m w_m \left| x_{im} - x_{jm} \right|^2},$$

where the summation is over the dimensions and w_m represents the attention weight on dimension m. Attention weights are restricted so that all must be greater than 0 and the sum of all attention weights is 1. The x_{im} is the value of exemplar i on dimension m. This value can be found through multidimensional scaling (MDS) studies or can be derived from the physical values used to generate the stimuli. In ACT-R, similarity is calculated through a city-block distance metric. Here, the difference between chunk i and pattern j is

$$M = \sum_m \left| x_{im} - x_{jm} \right|.$$

(Mismatch, Equation 6)

Although the equation above does not explicitly represent the weightings w_m from the EBRW equation, they are implicit in the scaling of the x_{im}. In both systems, the simi-

Table 1
Traces of the Various Methods of Classification in the ACT-R Model

1. Classification by guessing	
Cycle 10	Time 2.297: Guess
Cycle 11	Time 3.246: Random-Guess-Was-Right
Cycle 12	Time 3.296: Encode-4features
Cycle 13	Time 3.346: Study-Complete
Cycle 14	Time 3.396: Done-Right
2. Classification by retrieving examples	
Cycle 2173	Time 287.301: Choose-To-Classify-By-Exemplar-2feature
Cycle 2174	Time 287.479: Recall-2feature
Cycle 2175	Time 287.579: Recall-2feature
Cycle 2176	Time 287.679: Recall-2feature
Cycle 2177	Time 287.779: Recall-2feature
Cycle 2178	Time 287.879: Recall-2feature
Cycle 2179	Time 287.979: Recall-2feature
Cycle 2180	Time 288.079: Done-Classifying-By-Exemplar
Cycle 2181	Time 288.259: Correct-Finish-From-Exemplar
Cycle 2182	Time 288.309: Encode-2features
Cycle 2183	Time 288.359: Study-Complete
Cycle 2184	Time 288.409: Done-Right
3A. Classification by applying rule	
Cycle 1483	Time 348.160: Choose-To-Classify-By-Rule
Cycle 1484	Time 348.216: General-Rule-Match
Cycle 1485	Time 349.180: Feature1-Is-Nil
Cycle 1486	Time 349.233: Feature2-Is-Nil
Cycle 1487	Time 349.291: Feature3-Against-Rule-V1
Cycle 1488	Time 349.423: Feature4-Is-Nil
Cycle 1489	Time 349.474: Done-Applying-Presentation
Cycle 1490	Time 349.524: Done-Classifying-By-Rule
Cycle 1491	Time 350.183: Classification-By-Rule-Is-Right
Cycle 1492	Time 350.233: Increment-Correct-Count
Cycle 1493	Time 350.306: Imperfect-Rule-Satisfies-Stricter-Criterion
Cycle 1494	Time 350.365: Encode-4features
Cycle 1495	Time 350.415: Study-Complete
Cycle 1496	Time 350.465: Done-Right
3B. Classification by exception to rule	
Cycle 1497	Time 350.515: Choose-To-Classify-By-Rule
Cycle 1498	Time 350.568: Use-exception-4dim
Cycle 1499	Time 350.806: Classification-By-Rule-Is-Right
Cycle 1500	Time 350.856: Increment-Correct-Count
Cycle 1501	Time 350.911: Imperfect-Rule-Satisfies-Stricter-Criterion
Cycle 1502	Time 350.963: Encode-4features
Cycle 1503	Time 351.013: Study-Complete
Cycle 1504	Time 351.063: Done-Right

larity between a presented stimulus and a stored exemplar affects the probability of recalling a stored exemplar and the time to recall that exemplar. With the exception of this difference in metric, the combination of ACT-R Equations 2, 3, and 6 essentially is identical to EBRW. Note that the ACT-R equations were not at all fashioned to fit the categorization literature. Thus, the equivalence of ACT-R and EBRW reflects a significant convergence of theories developed to fit very different data sets.

The structure of declarative memory in ACT-R is also different than the EBRW model. In the EBRW model, multiple instances of an exemplar can be stored and can race against each other to be retrieved. In ACT-R, identical chunks are merged, so there can be only one copy of a stored exemplar. However, the strength of the merged chunk grows, according to Equation 3, since each merge contributes to the count N_i.[5]

Part 2 of Table 1 illustrates the sequence of production rule firings in a successful classification by an exemplar. The first production chooses the exemplar method, and then the next six productions implement the random walk. The threshold for the random walk in this example is 4, and six examples are retrieved to exceed this threshold—five voting in one direction and one in the opposite direction. Then, a production terminates the search, and another production processes the feedback. Finally, the snippet ends with the same final three productions as in Part 1 (except that, in this case, it is only a two-dimensional stimulus) to encode the current example.

Implementing RULEX in ACT-R

In the ACT-R implementation of the RULEX model, it is important to understand that classification rules are represented as chunks in declarative memory. That is to say,

a rule used for classification is not a production rule. The procedural/declarative distinction in ACT-R is made according to whether knowledge is explicit or implicit; declarative knowledge is explicit, whereas procedural knowledge is implicit. Therefore, it makes sense to represent rules for categorization as chunks in declarative memory because RULEX implies that such rules can be explicitly reasoned about by participants.

In the ACT-R implementation of the RULEX model, production rules implement both the search through alternative categorization rules and the assessment of the accuracy of particular rules. When the system attempts to find a new rule, a production fires, to select the kind of classification rule to look for. Imperfect rules are selected only if all the perfect rules have been exhausted. The process of forming exceptions to rules according to the RULEX specification is quite complex to implement in ACT-R, because of all the special cases (although it is by no means impossible, and we did successfully implement it as an exercise). In light of its complexity, we decided to replace it with a simpler system, in which exceptions are specified on all dimensions, unlike the partial specifications in RULEX.

Limitations on memory capacity are realized differently in the original RULEX and its ACT-R implementation. In RULEX, a parameter affects the rate at which new exceptions can be added to memory. In ACT-R, the subsymbolic mechanisms of declarative memory achieve similar effects. Consider the situation in which a permanent rule has been formed and many exceptions are already stored in memory but the system attempts to form a new exception. In RULEX, the limit on memory could prevent this new exception from being stored successfully. In ACT-R, the exception is guaranteed to be stored, but there is no guarantee that it will be available in a later recall. If the activation of the newly formed exception falls below the threshold for retrieval, it cannot be used. Therefore, the ACT-R implementation of RULEX maintains the assumption that memory is limited but does not enforce that assumption in the same way.

Parts 3A and 3B of Table 1 give traces of the production system's firing when it successfully classifies stimuli by the rule module. The difference between the two subparts is that in 3A it applies the general rule and in 3B it applies an exception. The first production to fire, Choose-To-Classify-By-Rule, retrieves the currently operative rule from declarative memory and chooses to apply it. If no exception can be retrieved (Part 3A), General-Rule-Match will start the process of comparing features. The model goes through the four dimensions, comparing them against the rule. The rule in this case is a one-dimensional rule, where only Dimension 3 is operative. In all the data sets we will be modeling in this paper, the actual rules are one-dimensional. When the feedback is received that the rule is right, Increment-Correct-Count updates the count required by RULEX for judging the viability of the rule, and Imperfect-Rule-Satisfies-Stricter-Criterion notes that the rule still satisfies the stricter criterion. In both Parts 3A and 3B, the model goes through the same final encoding of the example as in Parts 1 and 2.

Summary of the ACT-R Implementation of the EBRW and RULEX Models

As Table 1 illustrates, ACT-R implements its choice among the three methods essentially by a "big switch" that chooses one of the methods in the first production that fires in each snippet. Provided there are exemplars or rules that can be retrieved, the model will not choose the guess method, as in Part 1 of the table. The rule and exemplar methods will compete according to their relative success.

The mapping of the EBRW and RULEX models into ACT-R was fairly direct. However, it is not the case that the processes of ACT-R exactly correspond to the processes of EBRW or RULEX. The following highlights some of the differences.

1. ACT-R strengthens merged traces, whereas EBRW forms multiple traces that race against each other.

2. ACT-R uses a city block similarity metric, whereas the EBRW model typically uses an Euclidean metric. However, it should be noted that EBRW is not constrained to use an Euclidean metric.

3. ACT-R implements memory failures by retrieval limitations, whereas RULEX implements memory failures by storage limitations.

In each of these cases, ACT-R had a prior architectural commitment that forced us to take a somewhat different path. However, we did not think that these differences were critical, and they were not.

In addition to these differences, there is another category of issues involved in implementing the two models in the same ACT-R architecture. Generally, there is the question of whether the parameters that work for one model will work for another model. More specifically, there is the issue of whether the system of declarative memory that selects among traces in EBRW is consistent with the system that enforces memory limitations in RULEX. Also, there is the question of whether ACT-R would select among these two strategies to deliver the right mixture for a particular experiment. Therefore, this implementation effort is a nontrivial test both of the architectural compatibility of EBRW and RULEX and of the ACT-R architecture itself.

The key claim is that ACT-R has the facility to implement the essence of the two models and can predict when one model will be deployed versus the other. To put this claim to test, the hybrid ACT-R model was tested against three data sets. The first two data sets (Nosofsky & Palmeri, 1997b; Nosofsky et al., 1994) came from the papers that proposed the models that form the subsystems in the hybrid model. The third data set (Erickson & Kruschke, 1998) came from a study that showed the interaction of rule-based and exemplar-based systems. In describing the simulations of the individual data sets, we will try to focus on the most significant aspects of the models. The actual simulations are available for inspection from the *Published Models* link at the ACT-R home page (http://act.psy. cmu.edu/).

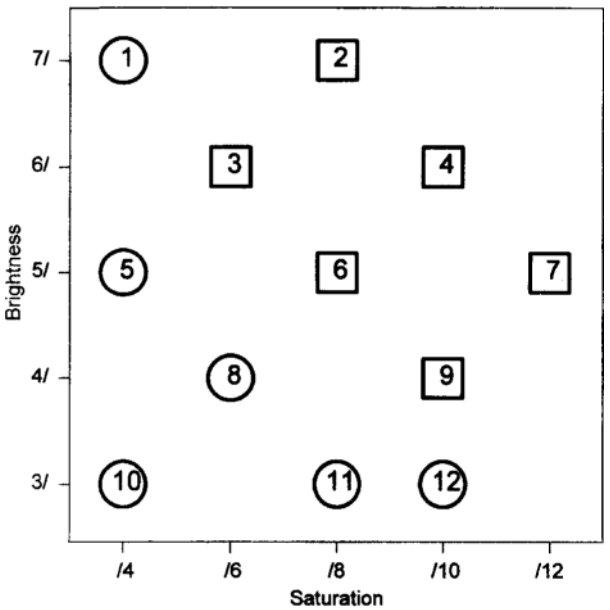

Figure 1. Schematic illustration of the color stimuli used in Nosofsky and Palmeri (1997b). Circles represent one category, and squares represent the other category. From "An Exemplar-Based Random Walk Model of Speeded Classification," by R. M. Nosofsky and T. J. Palmeri, 1997, *Psychological Review*, *104*, p. 274. Copyright 1997 by the American Psychological Association. Reprinted with permission.

DATA SET 1
Nosofsky and Palmeri (1997b)
Experiment 1

Nosofsky and Palmeri (1997b) report a study that tested their EBRW model. Three participants were presented with a set of 12 color squares. The stimuli had a constant hue but varied on the dimensions of brightness and saturation. The category structure of the stimuli is shown in Figure 1. The participants made judgments on these stimuli over 150 blocks. In each block, all 12 stimuli were presented for classification. The results showed power-law speed-up in classification for each participant and faster response times on stimuli that were further from the category boundary, as predicted by the EBRW model. An important feature of these stimuli for the purposes of the ACT-R simulation is that the dimensions are integral (Garner, 1974) and, so, participants cannot articulate the separate dimensions. As a consequence, when the RULEX submodel of ACT-R tries to formulate rules for classifying these stimuli, it will always experience failure. Thus, it will quickly switch to an exemplar approach.

A major component in modeling these data in ACT-R was setting the similarity values between differing levels of brightness and saturation appropriately. The original EBRW experiment included a posttest portion in which participants made similarity judgments about each possible pairing of stimuli. Nosofsky and Palmeri (1997b) used these judgments to derive an MDS solution for each participant. We used these solutions to set the subjective similarities for ACT-R among the various brightnesses and saturations. The MDS solution tells us the relative differences between the values, but it does not tell us the absolute differences. Nosofsky and Palmeri (1997b) estimated scaling parameters to convert these relative differences into absolute differences, and we did the same.

We estimated, for each participant, a retrieval time for each item retrieved, an *intercept* time corresponding to the time to encode the stimulus and respond, and a retrieval threshold for the activation of a chunk in order for it to be retrieved. There was also a counter threshold in the random walk for classifying a stimulus, but we held this constant at 4 across participants, in contrast to Nosofsky and Palmeri (1997b), who estimated different parameters for different participants (4 was their median value). These are given in Table 2, which gives the parameters for all the data sets. In addition, note that two parameters, the utility noise ($t_E = 2.2$) and the activation noise ($t_A = 0.78$), were set once for all simulations.

Table 2
Parameter Sets and Fits

Data Set	Retrieval Time Parameter (sec)	Retrieval Threshold Parameter, τ	Intercept Parameter (sec)	Counter-Threshold Parameter	Learning Correlations		Generalization Correlations	
					ACT-R	EBRW	ACT-R	EBRW
1. Nosofsky and Palmeri (1997b)								
Participant 1	0.10	−0.3	0.18	4	.94	.94	.87	.89
Participant 2	0.05 0.5	0.18 4	.77	.78 .97	.99			
Participant 3	0.05 0.0	0.25 4	.95	.96 .93	.95			
2. Nosofsky, Palmeri, and McKinley (1994)	0.05	0.8	0.20	1	–	–	.85	.92
3. Erickson and Kruschke (1998)	0.05 0.0	0.20 1	.99	.98 .94	.91			

Note—Retrieval time is the minimum step time, in seconds, for one retrieval in the random walk. Retrieval threshold is the minimum activation required for a chunk to be retrieved. Intercept is encoding and response time. Counter threshold is the magnitude of the counter value that triggers a decision in the random walk. The ACT-R columns show the correlations between the predictions of the ACT-R hybrid model and the data. The EBRW columns show the correlations between the predictions of the EBRW model and the data.

The fits that Nosofsky and Palmeri (1997b), report for their model came from searching for the best-fitting set of six parameters. We informally tried to find parameters that would give "close" values in a Monte Carlo simulation. In doing this, we adjusted the retrieval time, the intercept, and the retrieval threshold. Thus, ours might be viewed as a three- or four- (depending on how one views the constant counter threshold) parameter model, but not one that is optimized to give best fit.[6] Our goal is not to produce a better model, but just to establish that ACT-R can yield predictions that are similar to EBRW.

The learning data from this experiment are presented in Figure 2. These data show an approximate power-law decrease in response time over the training period. Table 2 also gives the correlations between predicted and observed data with the Nosofsky and Palmeri (1997b) predictions. Although the ACT-R fits are good and nearly identical to those from Nosofsky and Palmeri (1997b), it is worth noting that the learning in the early part of the curve is more complex in ACT-R, because the model is sometimes trying to form rules and failing (because the dimensions are not analyzable) and because the model is sometimes failing to retrieve exemplars and guessing. Both of these failed paths tend to lead to long reaction times. However, these events decrease in frequency with practice in a way that approximates power-law learning. Both Delaney, Reder, Staszewski, and Ritter (1998) and Rickard (1997) have shown that mixtures of strategies can yield approximate power-law learning.

Figure 3 shows the data from the last 120 blocks in terms of the time to classify the 12 stimuli in Figure 1. By this time, all the responses are based on the counting process, and the EBRW and ACT-R models basically correspond in terms of mean number of steps to classify each stimulus. Table 2 also shows the mean generalization correlations for ACT-R and the Nosofsky and Palmeri (1997b) model. Although the ACT-R fits are not quite as good as the Nosofsky and Palmeri (1997b) fits, they are clearly quite similar, and the ACT-R predictions have not gone through the same optimization process. The high correlation between the EBRW and the ACT-R models illustrates our earlier observation about the essential convergence between the two theories.

There are two features of the ACT-R architecture that are critical to the model's ability to implement essential aspects of the EBRW algorithm. First, the trace strengthening processes (Activation, Equation 2) and the latency function (Retrieval Time, Equation 5) are primarily responsible for the power-function speed-up (Figure 2). ACT-R's strengthening component has been noted to be equivalent to Logan's race among instances (e.g., Anderson, Fincham, & Douglass, 1999). Second, the partial matching process (Activation, Equation 2) and the stochastic noise in activations (Chunk Choice, Equation 3) combine to produce the different speed in classifying different stimuli (Figure 3). The present endeavor indicates that ACT-R partial matching is nearly equivalent to GCM's similarity-based retrieval. These two components and the random walk algorithm are the critical pieces to the EBRW account of the data. We just implemented the random walk, and this is not a test of ACT-R (except as noted below). However, the success of this effort supports the learning and partial matching processes of ACT-R. Moreover, since these components have participated in accounts of many other cognitive tasks (Anderson & Lebiere, 1998), the success of this effort indicates that the EBRW model is consistent with some general aspects of cognition.

With respect to the random walk algorithm, it is not a trivial matter that the timing parameters worked out, since ACT-R places definite limits on the range of times. Every production cycle takes at least 50 msec but retrievals from declarative memory can make the cycle take longer. On the other hand, we have argued (Anderson et al., 1998) that every production cycle cannot take much longer than 500 msec. Thus, in ACT-R, the timing of the steps are bounded to within an order of magnitude. By the end of the experiment, ACT-R was taking from about 50–100 msec to consider each exemplar in the random walk. We had worried that ACT-R would not be able to perform the steps in the random walk fast enough to match the data and were somewhat surprised that we were able to fit the data with the timing parameters in Table 2.

Although the ACT-R model is a fairly faithful implementation of the EBRW random walk algorithm, the ACT-R partial matching and strengthening mechanisms, which implement the algorithm, are different from the EBRW mechanisms, even if equivalent for the purposes of this experiment. One could imagine doing tests of the differences, focusing on things like the difference between the city block metric and the Euclidean metric. Although such tests would be valuable and might lead us to reformulate certain aspects of ACT-R theory, they would not change the conclusion that the two systems are nearly equivalent in the effects of their architectural assumptions.

Note that the RULEX component of this model did not play any role, because we did not provide the model with any rules to try. This corresponds to the observation that participants find it hard to articulate dimensional rules for the stimuli of the experiment. We do not claim that ACT-R provides any explanation of the inability of participants to analyze such color stimuli into their underlying dimensions. It simply represents that fact in its encoding of the situation. A basic premise of the model is that the rule component is available if and only if participants can identify the dimensions of the stimuli. This makes our interpretation of the RULEX system like the explicit verbal system of Ashby et al. (1998). We regard the cognitive neuroscience data that Ashby et al. cite for it as evidence for this component.

DATA SET 2
Nosofsky et al. (1994) Experiment 1

We will now consider an experiment reported by Nosofsky et al. (1994) that introduced the RULEX model. In this experiment, 227 participants were presented with 16 training blocks of nine trials each, showing line drawings of

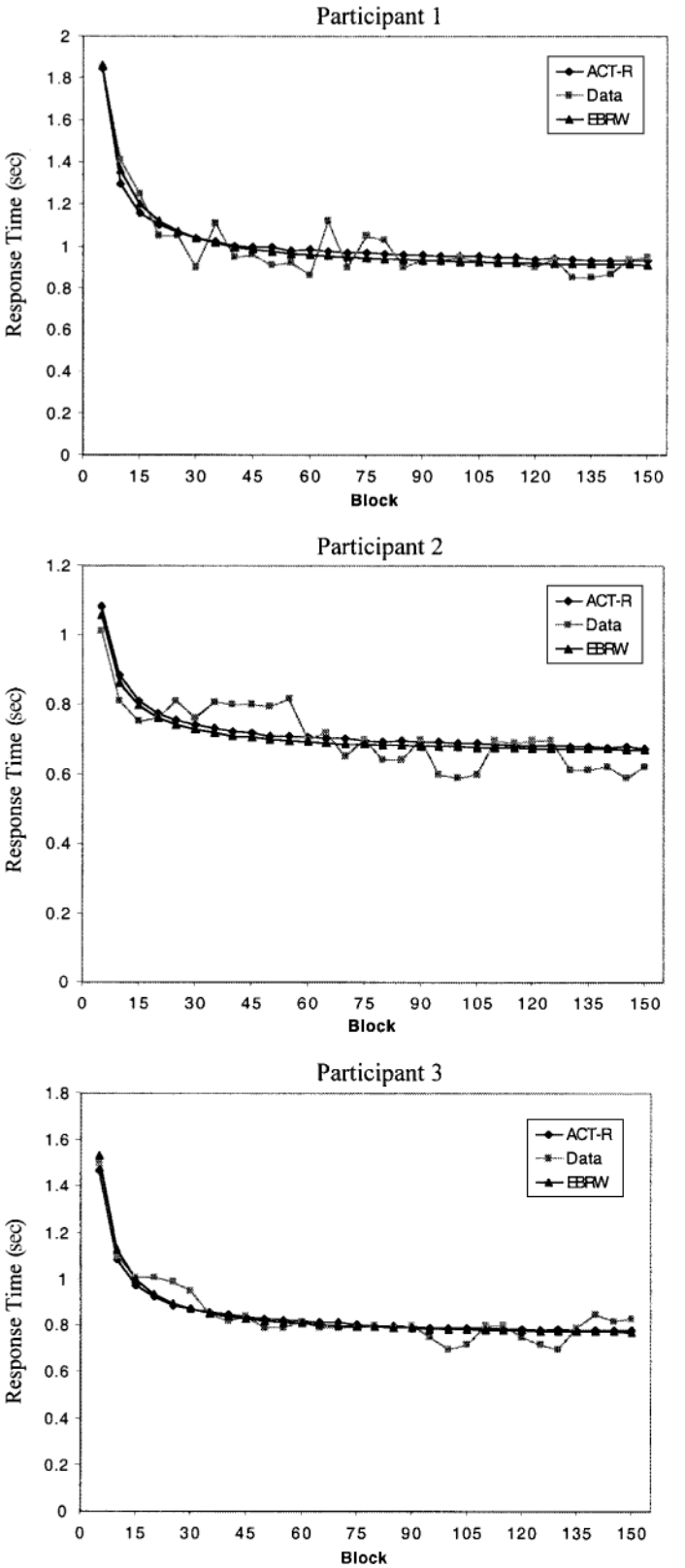

Figure 2. Comparison of ACT-R learning data with participant learning data from Nosofsky and Palmeri (1997b).

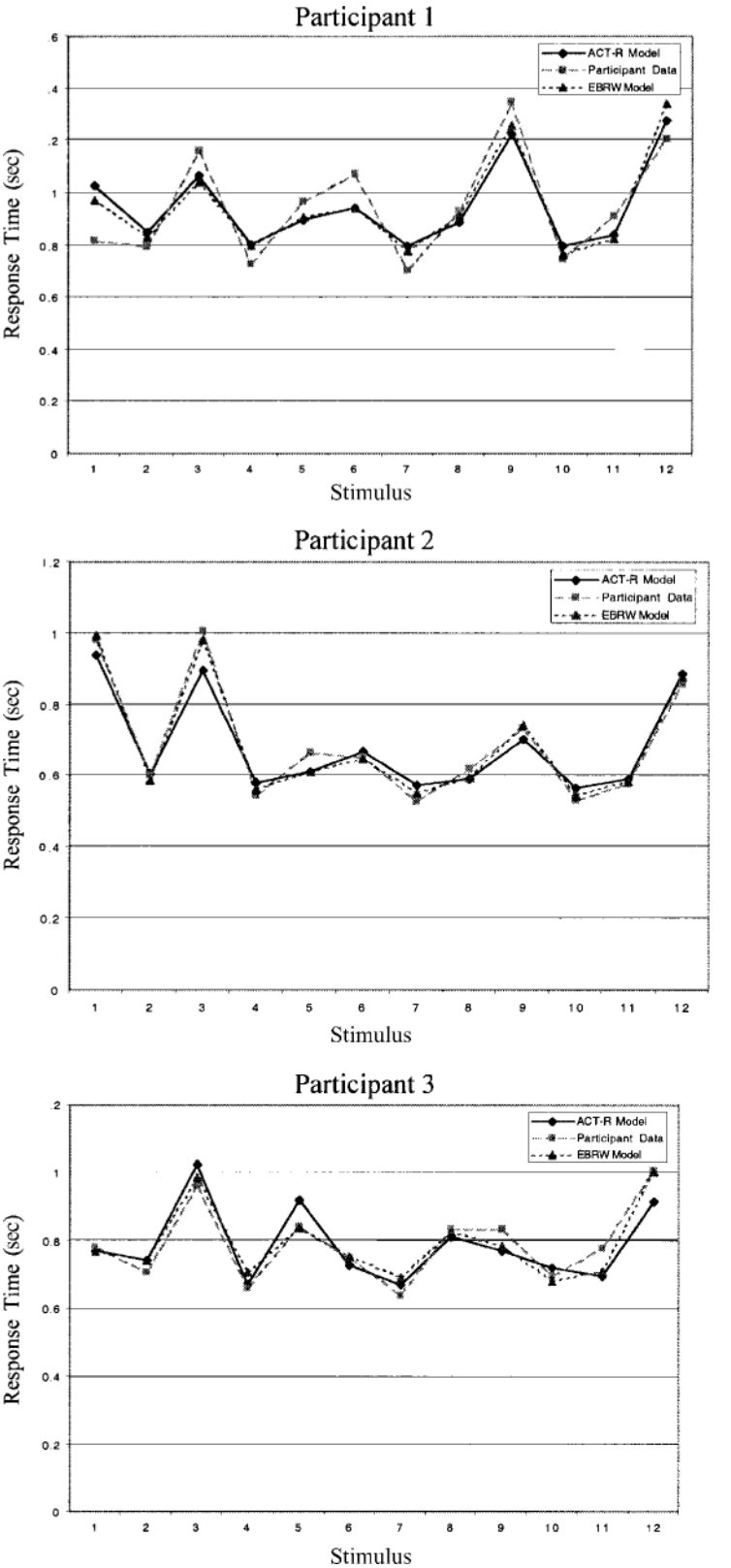

Figure 3. Comparison of mean response times for each individual stimulus from each participant and the ACT-R and EBRW model predictions for each participant from Nosofsky and Palmeri (1997b).

Table 3
Stimuli from Nosofsky, Palmeri, and McKinley (1994), Experiment 1

Category A	Category B	Transfer
A1: 1112	B1: 1122	T1: 1221
A2: 1212	B2: 2112	T2: 1222
A3: 1211	B3: 2221	T3: 1111
A4: 1121	B4: 2222	T4: 2212
A5: 2111		T5: 2121
		T6: 2211
		T7: 2122

rocket ships that varied on four binary-valued dimensions. During training, feedback was provided after each classification decision. After training, the remaining possible patterns from the stimulus space were presented, and the participants were prompted to make a classification decision on these novel stimuli. The complete category structure of the stimuli is presented in Table 2. These stimuli are isomorphic to those used in Experiment 2 of Medin and Schaffer (1978). No single-dimension rule or conjunctive rule can correctly classify all the stimuli in this experiment. Therefore, any successful rule-oriented classification strategy will require storing exceptions. There is enough structure in the stimulus set and the structure is simple enough that a rule-based approach does enjoy a fair amount of initial success. Therefore, it is possible to learn a rule that correctly classifies most of the stimuli and then learn a small number of specific exceptions to that rule.

The generalization patterns shown by participants on novel stimuli are given in Figure 4. There are seven transfer stimuli, each with two possible responses, for a total of 2^7 possible generalization patterns. Palmeri and Johansen (1999) suggest ignoring Transfer Stimuli 3 and 7 in Table 3, since they received the same classification on all bases. This reduces the 128 transfer patterns to $2^5 = 32$, which are graphed in Figure 4. Three blocks of transfer stimuli were presented, so a participant was said to classify a transfer stimulus into a given category if he or she responded with that category to the given stimulus on at least two of the three presentations. The two most common generalizations shown were AABBB and BBABA, which correspond to rules on Dimensions 1 and 3.

Nosofsky et al. (1994) fit a five-parameter version of RULEX to these data. One of these parameters was the strict criterion for single-dimension rules, which varied uniformly between .65 and .85, and we similarly allowed this criterion to vary in ACT-R. We set the difference between the binary values to be worth an M_i value of 2.25 (see Activation, Equation 2). The other ACT-R parameters are shown in Table 2. Note that we set the counter for the random walk to 1, thus classifying the stimulus on the first retrieved item, which effectively eliminates the random walk. Since there is no latency data reported in this experiment, there is nothing to be gained by the random walk process, and it can considerably lengthen the simulations.[7] The ACT-R simulation was run through the same 16 training blocks as the participants and given the same transfer trials. Our results are averages of 1,000 simulated participants.

The results of RULEX and the ACT-R model are also presented in Figure 4. These results show that the ACT-R hybrid model reproduces many of the major aspects of the participant data. The AABBB and BBABA generalization patterns are those most frequently generated by this model, which is in accord with the data. The first generalization pattern represents a rule based on Dimension 1, whereas the second generalization pattern represents a rule based on Dimension 3. The overall correlation with the data is .85 for ACT-R and .92 for RULEX. As in the case of Data Set 1, this establishes that the model can get in the range of the data without any special effort at tuning.

In any given run of the model, either the rule-based or the exemplar-based system can learn the category structure of the experiment. Over trials, the model tends to shift from trying rules to using examples. Figure 5 plots the proportion of example use as a function of trials (rule use never succeeds for Data Set 1, and so there is not a comparable plot). There are two reasons for the increased use of exemplars. First, should a high criterion be selected for rule success, no rule will exceed the criterion, and ACT-R will switch to examples.[8] Second, as the amount of experience increases, the examples become more and more strongly encoded, and retrieval becomes a faster way of classifying examples without sacrificing accuracy. As this is discovered, the conflict resolution (see Equation 1) changes its preference from rules to examples. Basically, ACT-R has realized a variation on Logan's (1988) exemplar model. The system starts working with inferred rules and switches to examples. As Figure 5 illustrates, this same tendency occurs in the third data set. The fact that nearly 50% of the classifications are example based by the end of the experiment explains why more of the classifications in Figure 4 are not of the form AABBB or BBABA. The dominance of these two transfer patterns is a consequence of the RULEX algorithm implemented in ACT-R, and not of the architecture. The RULEX algorithm by itself would produce over 30% choice of each of these patterns, rather than the approximately 15% displayed in Figure 4. The 30% is reduced to 15% because, in many of the runs, ACT-R has switched to exemplars. In the original RULEX model, the reduction in the frequency of these two generalization patterns occurs because of random slips and because sometimes an exception blocks a generalization. It is a prediction of the ACT-R architecture that the rule-based generalization will become more muted as participants practice the task more and switch more to example-based classification. This is not a prediction of the original RULEX model. Palmeri and Johansen (1999) report a decreased tendency to make rule-based classifications in longer versions of this experiment. J. D. Smith and Minda (1998) report a similar effect.

DATA SET 3
Erickson and Kruschke (1998) Experiment 2

Erickson and Kruschke (1998) described an experiment in conjunction with a connectionist hybrid model for category learning. They emphasized the effect of interaction

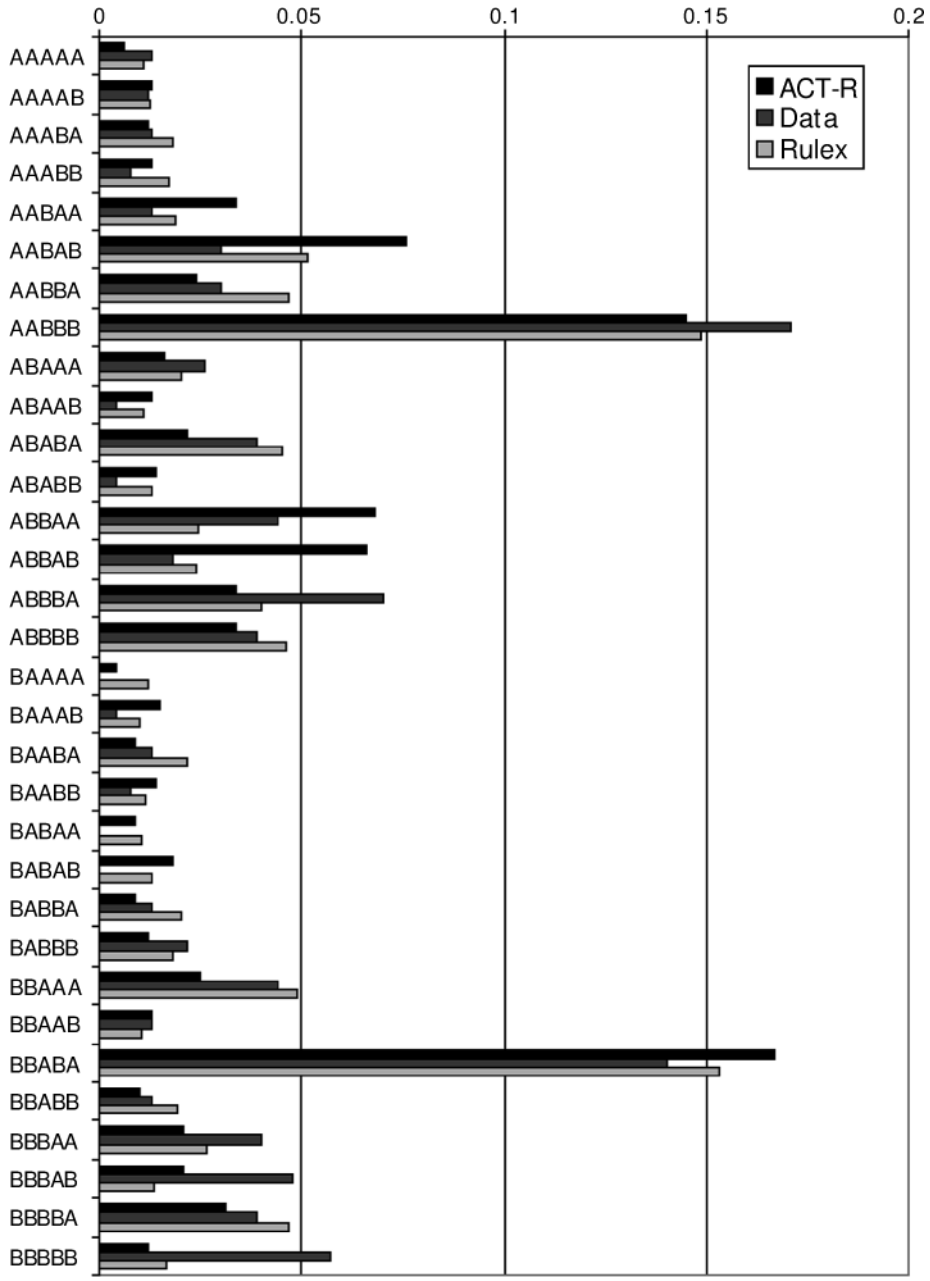

Figure 4. Comparison of generalization patterns shown in Nosofsky, Palmeri, and McKinley (1994) and by the ACT-R hybrid model. The categories are the 32 categories obtained by excluding responses to Transfer Stimuli 3 and 7 in Table 2.

between exemplar-based and rule-based modules of their hybrid categorization model. This interaction was studied by varying the frequency with which various stimuli were presented. Participants were presented with rectangles that varied in height and the location of the vertical line segment that the rectangles contained. These two dimensions formed a stimulus space that is illustrated in Figure 6. The height of the rectangle was the primary dimension; a rule that divided the stimulus space on the basis of rectangle height correctly classified all but two of the stimulus patterns. These patterns were exceptions, and each belonged to its own category (thus, there were four categories in all). Each dimension could take a discrete value from 0 to 9, and the patterns were presented on a display, with axes to indicate the numeric value of both dimensions.[9] All but four stimuli were presented once per block of training. One exception-classified stimulus (E2) and one rule-classified stimulus (R2) were presented twice per block. One

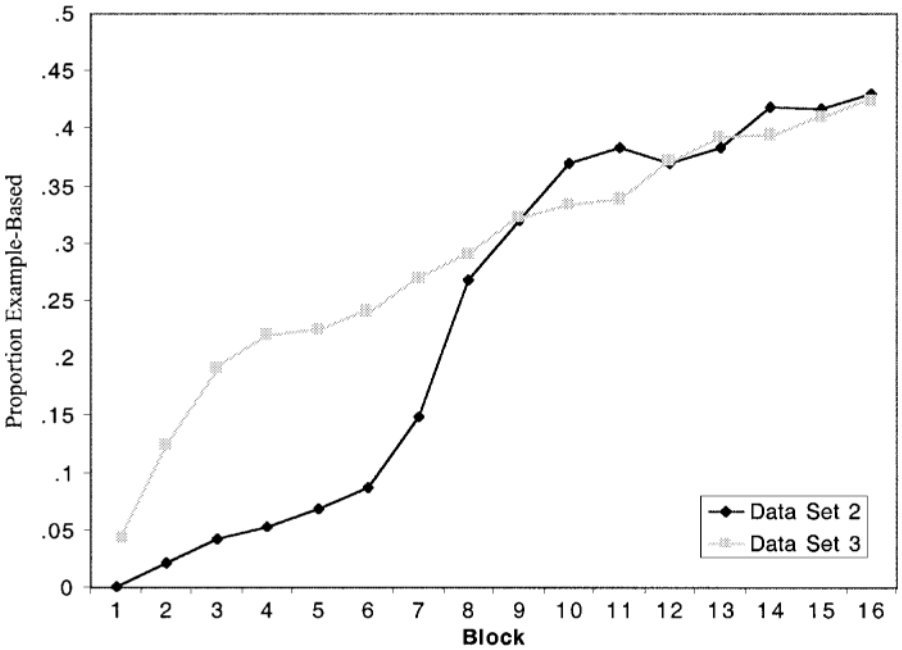

Figure 5. Percentage of classification by examples (in contrast to rules) as a function of trials for Data Set 2 (Nosofsky, Palmeri, & McKinley, 1994) and Data Set 3 (Erickson & Kruschke, 1998).

exception-classified stimulus (E4) and one rule-classified stimulus (R4) were presented four times per block. In transfer blocks, they measured the percentage of rule-appropriate responses to the eight patterns adjacent in the stimulus layout to each of these higher frequency stimuli.

Participants alternated between 16 blocks of training in which they were given feedback on their classifications and 16 transfer blocks in which no feedback was given. Part A of Figure 7 displays the data from this experiment. With respect to the training data, there is an effect of frequency on both exception-classified stimuli and rule-classified stimuli. Transfer stimuli are considered correct if participants give rule-appropriate responses even if the stimuli surround an exception. The participants generated a higher percentage of rule-appropriate responses to stimuli surrounding R4 than to those stimuli surrounding R2. Furthermore, the participants generated fewer rule-appropriate responses to stimuli surrounding E4 than to stimuli surrounding E2. Erickson and Kruschke (1998) used these data to argue against a pure rule-based model, because they claimed that this would not predict a frequency effect for rule-classified stimuli. They also used these data to argue against a pure exemplar model, because they claimed it would not predict that transfer stimuli in the vicinity of the exception stimuli would be classified a majority of the time according to the rule. The model that they proposed with their data, ATRIUM, was a connectionist model that, on each presentation of a stimulus, made a categorization decision by using a rule-based subsystem, made another decision by using an exemplar-based subsystem, and combined these two judgments to reach a weighted final decision.

The ACT-R model that we developed for this task could use either exemplars or a rule that separated the stimuli by a value of 4.5 on the height dimension. This rule is like the linear decision boundary rules proposed by Nosofsky and Palmeri (1998) in their extension of RULEX to continuous dimensions. Because the ACT-R hybrid model only tries one method (rule or exemplar) on a trial, rather than merging the results of two methods, it was not initially apparent to us that it could predict an effect of frequency of presentation on probability of rule-appropriate decisions. The decision to use a rule-based or exemplar-based approach is determined by the overall success of these approaches, rather than by the success with respect to a particular stimulus. When we ran the ACT-R model on this task, we found, to our surprise, that it did a good job in predicting qualitative trends in the data, as is indicated in Part B of Figure 7. With respect to correlations (Table 2), it does a slightly better job of predicting certain aspects of the data than does ATRIUM, whose predictions are illustrated in Part C of Figure 7.[10] The ACT-R model tends to slightly overpredict performance—a problem we could have corrected by adding a slip parameter, but the complication did not seem worth it. (Interestingly, ATRIUM tends to slightly underpredict accuracy.) The ACT-R parameters that were set for this model were the same as those for the previous model (Data Set 2), except that we changed the retrieval threshold and had to scale the similarities separately for this experiment. We scaled the differences among stimuli so that each unit difference (on either axis) in Figure 6 was worth an M_i of 1.5 (see Activation, Equation 2). Thus, for instance, if the stimulus had a height of 4, the target height

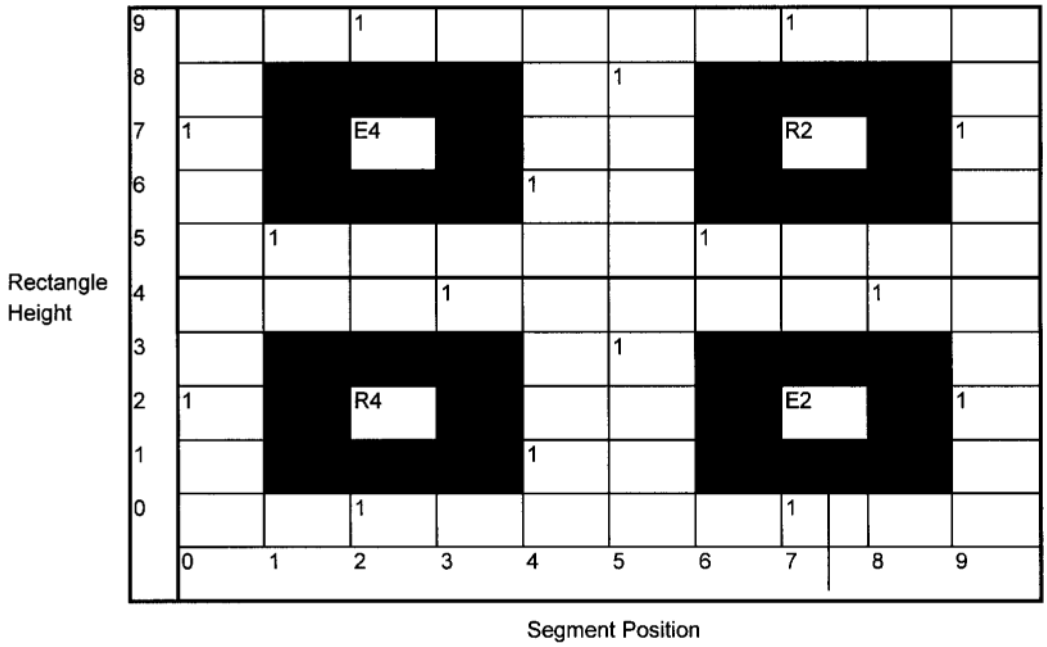

Figure 6. The category structure used by Erickson and Kruschke (1998). Each 1 indicates a stimulus that was presented once per training block. The cells labeled R2 and R4 indicate stimuli that could be correctly classified by the rule and were presented two and four times, respectively, per training block. The cells labeled E2 and E4 indicate stimuli that were not correctly classified by the rule and were presented two and four times, respectively, per training block. The shaded area indicates the transfer stimuli for which percentage of rule-appropriate responses were measured.

was 7, and the other dimension matched, the mismatch would be $3 * 1.5 + 0 \times 1.5 = 4.5$.

Figure 5 displays the growth in exemplar use over trials. The ACT-R predictions depend on the fact that it is using a mixture of exemplars and rules. The reason it tends to classify transfer stimuli according to a rule, even when they are close to an exception, is that a majority of the trials are rule based even at the end of the experiment. ACT-R tends to get the exceptions correct because they are stored as exceptions in the RULEX portion of the model. Note that these exceptions are represented twice—once as exceptions to the rules and once as exemplars. The frequency effects largely come from the trials in which exemplars are used.

Erickson and Kruschke (1998) thought that their data provided evidence for something they called *representational attention.* This was the ability to focus on the exemplar module for exceptions and the rule module for the other stimuli. This was the way they were able to achieve high accuracy on exceptions and high accuracy on the other stimuli as well. In contrast to their ATRIUM model, our ACT-R model was able to achieve this high accuracy even though it chooses to use the rule module or exception module independent of the stimulus. It can maintain high accuracy because its RULEX component checks for exceptions. Erickson and Kruschke pointed out that RULEX by itself would not be able to produce the frequency effects in the data. This is what is produced by our exemplar component. Thus, our hybrid model needs both the exceptions stored in RULEX, to achieve accurate classification of the studied exceptions, and the exemplars, to produce a frequency effect in transfer.

GENERAL DISCUSSION

In this paper, we have taken two algorithms for classification, the random walk in EBRW and the rule search in RULEX, and implemented them with relatively few modifications in the ACT-R architecture. The fact that these algorithms could be implemented in ACT-R is not a trivial result. For instance, it would be hard to implement the elaborate sequential decision structure of RULEX in most connectionist architectures. Furthermore, the fact that we could implement the algorithms does not necessarily imply that they would have behaved in a way that matched up with the data. If we had implemented these algorithms in earlier versions of the ACT-R architecture (Anderson, 1976, 1983, 1993) or in other production system architectures (e.g., Just & Carpenter, 1992; Kieras & Meyer, 1997; Newell, 1991), they would have behaved differently. The successful performance of these algorithms in ACT-R required certain properties of the declarative and procedural components of the ACT-R architecture.

The declarative component of ACT-R (in the guise of the Activation Equation 2, the Chunk Choice Equation 3, the Retrieval Probability Equation 4, and the Retrieval Time

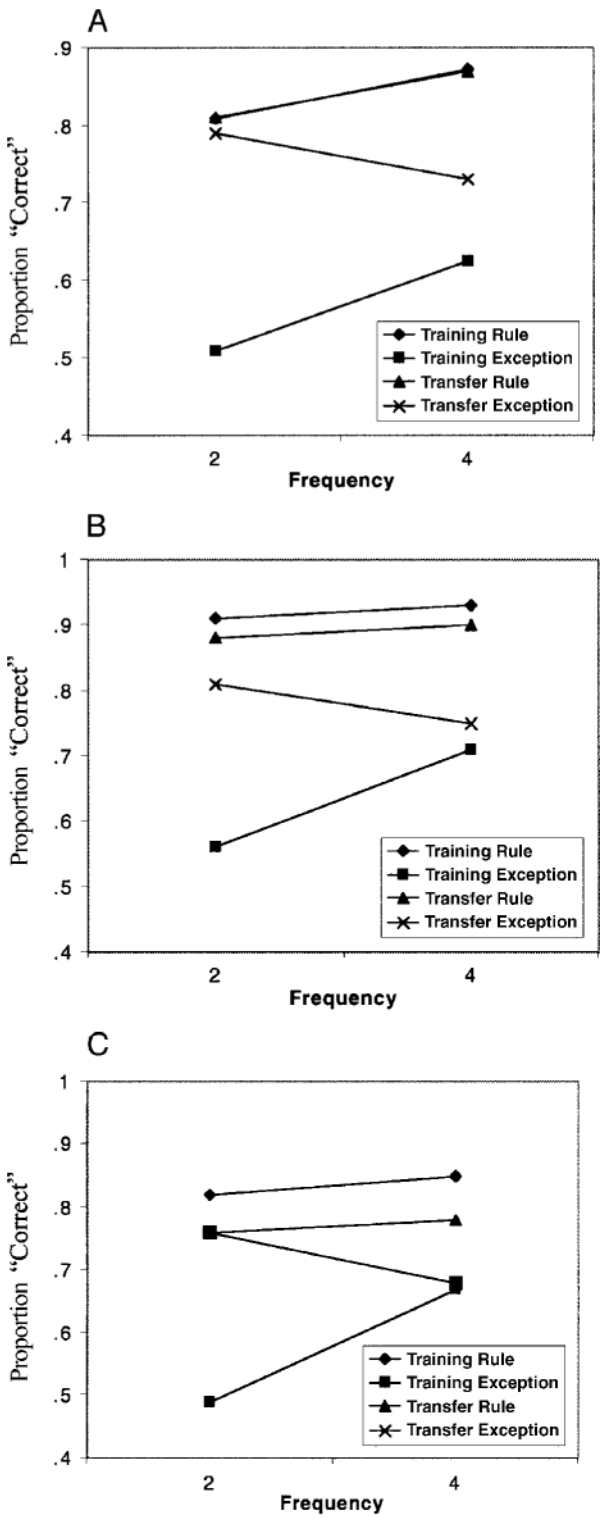

Figure 7. Proportion "correct" (rule-appropriate responses are scored correct for transfer stimuli) as a function of frequency. (A) Data from Erickson and Kruschke (1998); (B) ACT-R predictions; (C) ATRIUM predictions.

Equation 5) was able to produce the frequency effects in Data Sets 1 and 3. The partial matching process was responsible for the similarity profiles in Data Set 1 and the transfer performance in Data Set 3. The success of ACT-R's declarative component offers a significant generalization in three ways. First, it shows that ACT-R is another way of characterizing memory within the categorization algorithms. Second, it shows that the same memory characterization can work within both the EBRW and the RULEX algorithms. Third, to the extent that ACT-R has been applied to domains other than categorization, it shows that the same memory processes underlie categorization as other tasks.

The other contribution of this paper was to show how the two categorization algorithms can coexist together. That is to say, they can work in a single cognitive architecture in which they are constrained to share the same parameters (such as activation noise and decay). Moreover, they can coexist in an architecture that supports a wide variety of other cognitive processes. Here, the credit goes to ACT-R's procedural component. The procedural component (in the guise of the Conflict Resolution Equation 1) was responsible for the mix of exemplar- and rule-based strategies. Exemplar use dominated from the beginning for Data Set 1, because it was not possible to formulate rules. In contrast, for Data Sets 2 and 3, rule use dominated early and slowly gave way to exemplar use. The exemplar use played a role in the account of the second data set, and as Palmeri and Johansen (1999) have shown, it can be more significant with more practice. The mixture of rule and exemplar judgments was absolutely critical to our success in accounting for Data Set 3. The transition from rule-based to exemplar-based classification is rational and is captured by the conflict resolution process in ACT-R. Rule-based classification is more economical in terms of memory structures encoding the exemplars but is less efficient in terms of processing time. As the memory structures encoding the exemplars become strengthened, ACT-R transitions to exemplar-based classification.

It is worth emphasizing that both the exemplars and the rules are represented as declarative chunks in the current ACT-R model. Production rules basically "interpret" these chunks. This contrasts with a much earlier ACT proposal (Anderson, Kline, & Beasley, 1979) that represented both instances and abstractions as production rules. This production rule implementation of categorization has recently been extensively modified and elaborated by Vandierendonck (1995). We choose to implement categorical knowledge declaratively because we believe participants can describe their knowledge and only declarative knowledge can be described in ACT-R.

The suspicion is sometimes expressed that an architecture like ACT-R can fit any pattern of data. However, this is not true in general and is not true in this case. In comparison with other models, we are constrained not only to fit the data at hand, but to do it in a way that is consistent with our models for other domains. ACT-R is a system

that actually performs the task in real time and, moreover, has strong commitments to the time that each step of cognition takes. These commitments are in the form of parameters, and bounds have been established for these parameters in fitting other data sets in other domains. This makes the first data set from Nosofsky and Palmeri (1997b), with its timing information, perhaps the most demanding. As we noted, it was not a trivial matter that ACT-R was able to implement the EBRW algorithm. If the human data had involved steps in the random walk that were only half as long, ACT-R would not have been able to do the task in human time.

The strategic decision to implement EBRW and RULEX in ACT-R makes this effort more constrained than if we had fashioned our own exemplar- and rule-based model. Since the implementations of these models succeeded (a nontrivial result), our exemplar module or our rule-based module are separately no more able to fit any pattern of data than are the original models. However, one might argue that the combination of two modules is more flexible than either by itself. This is not true in this case, because of the constraints that the architecture brings. For instance, we are committed to the prediction of increased exemplar use in tasks in which these two strategies mix—a prediction that pure models do not make. Also, our hybrid model commits us to predicting the interaction observed in the Erickson and Kruschke (1998) data (Figure 7). All parameter adjustment could do would be to change the size of the main effects and interactions. Thus, the ACT-R model would have been disconfirmed had the results of these experiments been in the opposite direction. These constraints on the mixture of the strategies come from the basic declarative and procedural mechanisms in ACT-R.

Table 2 lists the parameters used to fit the individual data sets. In addition to these, we set two global parameters especially for these experiments. These are the noise parameters for utility ($t_E = 2.2$) and activation ($t_a = 0.78$) and they are in the range used for other tasks. Although ACT-R has other parameters, these were set at default values established in past research. The parameters in Table 2 were set to match the performance level in the experiments. Except for the 1st participant in Nosofsky and Palmeri (1997b), who displayed slower times, the retrieval time parameter was kept constant at 50 msec. Similarly, there was no serious effort to estimate the counter threshold parameter or the intercept parameter. It was the activation threshold parameter, τ, that varied substantially across data sets, producing the different levels of performance. Anderson et al. (1998) noted that this parameter also varied substantially in their fit to different list-learning experiments. It does appear to be one that captures the performance differences across experiments.

With respect to the latency structure of the data, the current ACT-R model assumes that all the dimensions of a stimulus are encoded at once, as do EBRW and RULEX. However, Lamberts (1998) has argued the various dimensions are encoded separately, and Lamberts and Freeman (1998) show, with stimuli like those from Data Set 2, that this is an important consideration in predicting latency. Clearly, this is a direction to proceed in elaborating the model presented here.

Nosofsky and Johansen (2000) have recently made the case that an exemplar model can account for all data that have been used to argue for rule-based processing. This includes the last two data sets[11] that we modeled in this paper. They do not claim that the exemplar model provides a superior account but, rather, question whether any existing data sets conclusively establish the need for something other than exemplars. Although one could strive to find the decisive experiment that decides the issue once and for all, perhaps a more promising approach is to try to see how categorization behavior fits in with a more complete characterization of human cognition. Constraints from other domains can point the way to the correct model of human categorization. This paper is a first step toward that goal.

EBRW is just one of many models for doing example-based classification, and RULEX is just one of many models for using abstractions to make category judgments. To what degree do these results provide support for the EBRW and RULEX algorithms specifically? With respect to EBRW, its essential feature is the random walk that allows similarity to influence retrieval time. We suspect that the number of steps in the random walk in EBRW are too many. Already in our fit to Data Sets 2 and 3 (where we were not constrained by an existing model), we set the decision bound to one step. For the first data set, where we had a decision bound of four steps, we noted that ACT-R, given its minimal cycle time, was barely able to complete the processing in human time. Nosofsky and Palmeri (1997a) report a simpler experiment in which they estimated that participants had decision bounds of six steps and made their decisions in under 500 msec. ACT-R could not reproduce this result.

With respect to RULEX, it seems that an essential feature of the algorithm is its ability to have exceptions override a general rule. As we discussed earlier, it would have been difficult to simulate the Erickson and Kruschke (1998) data without this feature. It also seems that the Nosofsky et al. (1994) data (as well as the earlier concept-learning literature) indicate that people will try to classify stimuli according to a single dimension. Although we think that RULEX captures some essential features, we were stuck by the complexity of implementing the rule-search algorithm in terms of doing an exhaustive search through rules and keeping track of lax and strict criteria. Perhaps some simpler system could work as well. For instance, the EPAM model described by Gobet, Richman, Staszewski, and Simon (1997) seems to have all the essential properties of RULEX.

Another question is whether rules are the only form of abstraction or whether there are other ways of abstracting category information. In particular, one can imagine that participants have prototypes, perhaps with information about their variance, as well as about their central tendency. Our earlier rational model of categorization (An-

derson, 1991) and the earlier Anderson et al. (1979) model both proposed that participants stored multiple prototype-like representations. It is of interest that J. D. Smith and Minda (1998) propose that participants mix a prototype strategy with an exemplar strategy, whereas Ashby et al. (1998) propose that participants mix a prototype with a rule-based strategy. This prototype system may be implemented in the implicit, striatal system that Ashby et al. argue for and Poldrack, Prabhakaran, Seger, and Gabrieli (1999) find fMRI evidence for. Perhaps the right conclusion is that participants can use all three of the strategies: prototypes, exemplars, and rules.

There has been recent discussion about the ability of cognitive neuroscience data to select among alternative proposals for categorization. For instance, Ashby et al. (1998) argue that such data can be decisive. They argue that the fact that amnesiac populations appear to categorize successfully (Knowlton & Squire, 1993; Squire & Knowlton, 1995) can be used to reject exemplar models, since these patients cannot remember the examples but can categorize them. On the other hand, Nosofksy and Zaki (1998) and Palmeri and Flanery (1999) have shown that these results can be explained within an exemplar framework. It seems unlikely that such data really rule out any categorization strategy altogether. It seems more likely that neuroimaging data, such as those of E. E. Smith et al. (1998), will help identify which strategies particular participants are using in particular experiments.

In summary, it is probably too strong to say that this exercise has uniquely supported EBRW or RULEX; rather, it indicates that these theories capture some important aspects of categorization behavior. Likewise, it would be too strong to conclude that this research has uniquely supported the ACT-R architecture. Rather, it has relied on two critical features of that architecture, which are the activation-based declarative memory and a procedural system in which different paths are chosen according to their relative utility. Perhaps the safest conclusion to make is that this kind of architecture can implement the kinds of strategies that participants use in categorization. This is significant, because this is the kind of architecture that has successfully modeled participant behavior in other domains.

REFERENCES

Ackley, D. H., Hinton, G. E., & Sejnowsky, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147-169.

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

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

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory & Language*, **38**, 341-380.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1120-1136.

Anderson, J. R., Kline, P. J., & Beasley, C. M. (1979). A general learning theory and its application to schema abstraction. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 277-318). New York: Academic Press.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, **107**, 442-481.

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary for new directions. *Mathematical Cognition*, **1**, 3-34.

Bower, G. H., & Trabasso, T. R. (1963). Reversals prior to solution in concept identification. *Journal of Experimental Psychology*, **66**, 409-418.

Campbell, J. I. D. (1997). On the relation between skilled performance of simple division and multiplication. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1140-1159.

Delaney, P., Reder, L. M., Staszewski, J., & Ritter, F. (1998). The strategy specific nature of improvement: The power law applies by strategy within task. *Psychological Science*, **9**, 1-7.

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

Erickson, M. A., & Kruschke, J. K. (in press). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*.

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **8**, 37-50.

Gobet, F., Richman, H., Staszewski, J., & Simon, H. A. (1997). Goals, representations, and strategies in a concept attainment task: The EPAM model. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 37, pp. 265-290). San Diego: Academic Press.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.

Hinton, G. E., & Sejnowsky, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & the PDP group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 282-317). Cambridge, MA: MIT Press.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs* (Whole No. 123).

Just, M. A., & Carpenter, P. N. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance and application to human–computer interaction. *Human–Computer Interaction*, **12**, 391-438.

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

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

Lamberts, K., & Freeman, R. P. J. (1998). Building object representations from parts: Tests of a stochastic sampling model. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 904-926.

Lebiere, C. (1998). *The dynamics of cognition: An ACT-R model of cognitive arithmetic.* (CMU Computer Science Dept. Technical Report CMU-CS-98-186). Pittsburgh: Carnegie Mellon University.

Levine, M. (1975). *A cognitive theory of learning.* Hillsdale, NJ: Erlbaum.

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

Lovett, M. C. (1998). Choice. In J. R. Anderson & C. Lebiere (Eds.), *Atomic components of thought* (pp. 255-296). Mahwah, NJ: Erlbaum.

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

Newell, A. (1991). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

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

NOSOFSKY, R. M., & ALFONSO-REESE, L. A. (1999). Effects of similarity and practice on speeded classification response times and accuracies: Further tests of an exemplar-retrieval model. *Memory & Cognition*, **27**, 78-93.

NOSOFSKY, R. M., & JOHANSEN, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychological Bulletin & Review*, **7**, 375-402.

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

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

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

NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.

NOSOFSKY, R. M., & ZAKI, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, **9**, 247-255.

PALMERI, T. J., & FLANERY, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, **10**, 526-530.

PALMERI, T. J., & JOHANSEN, M. K. (1999). Prototypes, rules, and instances in category learning. *Abstracts of the Psychonomic Society*, **4**, 98.

POLDRACK, R. A., PRABHAKARAN, V., SEGER, C. A., & GABRIELI, J. D. E. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, **13**, 564-574.

RAAIJMAKERS, J. G. W., & SHIFFRIN, R. M. (1981). Search of associative memory. *Psychological Review*, **88**, 93-134.

REDER, L. M. (1987). Strategy selection in questions answering. *Cognitive Psychology*, **19**, 90-138.

REDER, L. M. (1988). Strategic control of retrieval strategies. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 227-259). San Diego: Academic Press.

REDER, L. M., & RITTER, F. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 435-451.

REDER, L. M., & SCHUNN, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45-77). Mahwah, NJ: Erlbaum.

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

RICKARD, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, **126**, 288-311.

ROSCH, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-223.

SIEGLER, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, **117**, 258-275.

SMITH, E. E. (1989). Concepts and induction. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 501-526). Cambridge, MA: MIT Press.

SMITH, E. E., PATALANO, A. L., & JONIDES, J. (1998). Alternative strategies of categorization. *Cognition*, **65**, 167-196.

SMITH, J. D., & MINDA, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1436.

SMITH, J. D., MURRAY, M. J., & MINDA, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 659-680.

SQUIRE, L. R., & KNOWLTON, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, **92**, 12470-12474.

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

WIXTED, J. T., GHADISHA, H., & VERA, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 523-538.

NOTES

1. There is considerable discussion about the exact form of the speed-up and whether it conforms to a power function or some other function, such as an exponential function (e.g., Heathcote, Brown, & Mewhort, 2000). This is not an issue in the current paper.

2. These three strategies are very similar to the three in J. D. Smith and Minda (1998), except that J. D. Smith and Minda use a prototype abstraction strategy, rather than a rule strategy. This model also has some similarities with Reder and Schunn (1996) and Rickard (1997) in that it chooses among strategies, including retrieval and a rule-based procedure.

3. There are other constant terms to this equation (see Anderson & Lebiere, 1998, especially p. 124), but they effectively cancel out in the present applications.

4. Activation is like log familiarity in SAM (Gillund & Shiffrin, 1984; Raaijimakers & Shiffrin, 1981), and Equation 3 is the formal equivalent of the sampling probability in that equation.

5. In ACT-R, retrieving a chunk also increases the activation of that chunk. That increased activation means that the same chunk will have higher probability of recall in future steps of the random walk. In the early stages of an experiment, all chunks have relatively equal activations, but minor differences can be rapidly amplified through this positive feedback loop, leading to runaway strengthenings. To eliminate such runaway strengthenings in the ACT-R model, the activation of stored exemplars is held to a fixed rate of growth. Thus, in ACT-R, as in EBRW, there is one strengthening (or exemplar formed) each time a stimulus is presented, and the chunk representing the correct stimulus–category pairing is strengthened.

6. Since the results come from Monte Carlo simulations, there are not analytic equations that enable most parameter estimation procedures. Given the complexity of the simulations and the number of runs required to obtain stable estimates, it is not feasible to do an exhaustive search of the parameter space.

7. Since ACT-R's retrieval of instances is already stochastic, the random walk is not required to predict probability of choice.

8. Note that unlike the original RULEX, the ACT-R implementation tries only single-dimension rules and does not search through various two-dimensional classification rules.

9. Nosofsky and Johansen (2000) wondered about the explicit presentation of numeric values. They show, in a follow-up study, that participants would behave much like the exemplar model if this numeric value were removed. This is what our model would predict if the effect of removing the numeric values was to eliminate the ability to formulate explicit rules—see the discussion at the end of Data Set 1.

10. It should be stressed, however, that they fit a great deal more data than what is given in Figure 7.

11. Actually, they modeled the first experiment (rather than the second) from Erickson and Kruschke (1998). More recently, Erickson and Kruschke (in press) have produced results that cannot be so explained.

(Manuscript received April 11, 2000;
revision accepted for publication January 22, 2001.)