Psychonomic Bulletin & Review
2005, 12 (1), 24-42

The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models

MARK A. McDANIEL
Washington University, St. Louis, Missouri

and

JEROME R. BUSEMEYER
Indiana University, Bloomington, Indiana

The purpose of this article is to provide a foundation for a more formal, systematic, and integrative approach to function learning that parallels the existing progress in category learning. First, we note limitations of existing formal theories. Next, we develop several potential formal models of function learning, which include expansion of classic rule-based approaches and associative-based models. We specify for the first time psychologically based learning mechanisms for the rule models. We then present new, rigorous tests of these competing models that take into account order of difficulty for learning different function forms and extrapolation performance. Critically, detailed learning performance was also used to conduct the model evaluations. The results favor a hybrid model that combines associative learning of trained input–prediction pairs with a rule-based output response for extrapolation (EXAM).

Human concepts are complex and varied and serve a myriad of purposes. One way in which concepts are used is in learning how to categorize people or things and infer properties from category membership. Historically, this view of concepts has dominated the theoretical and empirical literature in cognitive psychology. But this view of concepts is too restrictive; another important way in which concepts are used is in learning *functional relationships* between continuous variables and in making predictions about one variable on the basis of another (Bourne, Ekstrand, & Dominowski, 1971; Uhl, 1963). There are many examples of such relationships that we encounter every day, such as predicting job performance on the basis of intelligence, anticipating mood level on the basis of stress intensity, forecasting interest rates on the basis of inflation rates, predicting harvest yields on the basis of amount of rainfall, and so on (Hammond, 1955; Hoffman, 1960). Learning functional relations between causes and effects is fundamental to the formation of intuitive theories about how the world works, and these predictions guide subsequent decisions about how to control the world (Hammond & Stewart, 2001; also see Murphy & Medin, 1985). For example, in order to control the economy we need to know how increases in interest rates affect consumer spending, which in turn affects manufacturing and employment rates.

The literature reflects an imbalance in the amount of attention devoted to categorization relative to function learning, with extensive progress in both empirical and theoretical understanding of categorization (Estes, 1994; Lamberts & Shanks, 1997), and much less empirical interest and theoretical progress in understanding function learning. Given the importance of function learning for human conceptual activity, the dearth of theoretical development in this area is a serious omission.

The purpose of this article is to provide the foundations for a more formal, systematic, and integrative approach to function learning that parallels the existing progress in category learning. This is accomplished by developing mature and complete models on the basis of preliminary ideas from the function-learning literature. We first provide an overview of the initial theoretical approaches and highlight their limitations. Next, we develop a number of formal models that provide a more comprehensive specification of function learning than the initial models have. Finally, we evaluate and contrast how well the models account for a range of basic learning and transfer findings.

Before presenting the theoretical approaches, we need to describe briefly the function-learning paradigm. The present article focuses on single input–output function-learning experiments in which a single cue x is mapped by a continuous function F into a single criterion z. In a typical experiment, participants are initially provided a neutral cover story that verbally describes the experimental task but provides little or no direct information

Preparation of this article was supported in part by National Institute of Mental Health Grant MH068346 to both authors. We thank Lyle Bourne, Stephan Lewandowsky, Todd Maddox, and one anonymous reviewer for helpful comments on an earlier version of this article. Correspondence concerning this article should be addressed to M. McDaniel, Department of Psychology, Washington University, One Brookings Drive, St. Louis, MO 63130-4899 (e-mail: mmcdanie@artsci.wustl.edu).

about the cue–criterion relation. This is followed by training with a sequence of cue–criterion training pairs. On each trial, the value of a predictor cue x is presented, and the participant's task is to predict the value of a criterion z. Immediately following the individual's prediction y, outcome feedback about the criterion z and the prediction error $e = (y - z)$ is provided. After a couple of hundred training trials, participants form some type of concept about the functional relation. Following the training phase, the participants' conceptual knowledge is tested during a *transfer* phase by presentation of novel cue values without feedback. There are two types of transfer tests: interpolation and extrapolation. An *interpolation test* is defined by presentation of a novel cue value that falls between two previously experienced training values. An *extrapolation test* is defined by presentation of a novel cue value that falls below or above all of the previously experienced training values.

INITIAL THEORETICAL FRAMEWORKS AND LIMITATIONS

For many years, the predominant general view regarding function learning has been the rule learning approach. According to this approach, the learner constructs abstract representations that summarize the ensemble of cue value–criterion (response) value pairings used to teach the function. Most frequently, polynomial rules have been proposed as learners' underlying representations of the mappings between cue values and response values (Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991), although other rules, such as Fourier rules, have also been suggested (Carroll, 1963). Polynomial rules have also been used in category learning to partition a multidimensional stimulus space into categories (Ashby & Maddox, 1998).

Experiments investigating the polynomial rule models have been directed at transfer tests, which were designed to examine interpolation and extrapolation performance. In several studies, a polynomial rule model has been evaluated on interpolation performance, and the model accounted well for human performance (DeLosh, Busemeyer, & McDaniel, 1997; Juslin, Olsson, & Olsson, 2003; Koh & Meyer, 1991). However, predictions of the polynomial rule model for extrapolation performance have been examined in only two of these studies. Juslin et al. found that a rule model accounted for extrapolation, whereas in a different paradigm DeLosh et al. found that the rule model grossly overestimated the extrapolation accuracy that human learners actually display, especially as cue values became farther from training values.

Busemeyer, Byun, DeLosh, and McDaniel (1997; DeLosh et al., 1997) considered an alternative model for function learning based on associative learning mechanisms. Similar to earlier category learning models (cf. Kruschke, 1992), this model assumes that the learner forms direct associations between each stimulus and the corresponding criterion and stores all of these individual associations without abstracting any summary information. Then, a response magnitude is retrieved whenever the associated cue value is presented.

The attractiveness of the associative model is that it emanates from a long tradition of learning theory in psychology and incorporates well-specified learning algorithms (Hinton & Anderson, 1981). This model fails, however, when faced with an extrapolation task based on cue values outside the training range. Under these conditions, the model grossly underestimates the amount of extrapolation that humans are willing to generate (Busemeyer, Byun, et al., 1997; DeLosh et al., 1997; Juslin et al., 2003). Accordingly, Busemeyer, Byun, et al. and DeLosh et al. endowed the associative learning model with an extrapolation rule response mechanism (extrapolation associative model, or EXAM). DeLosh et al. demonstrated that EXAM reproduced the human extrapolation performance much better than the polynomial rule model did.

In light of the significant failure to account for one set of extrapolation findings (DeLosh et al., 1997), some theorists have tended to discard the polynomial rule models for function learning (DeLosh et al., 1997; Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Kalish, & Ngang, 2002). However, this conclusion may be premature. Other researchers continue to argue in favor of the rule-based models (Juslin et al., 2003), and, furthermore, formal instantiation of rule-based models has been limited on several critical dimensions. The upshot is that (1) the learning data have been ignored in comparing rule-based versus associative-based models and (2) existing comparisons between these models on extrapolation are not definitive. The latter observation follows in part because conclusions concerning extrapolation are contingent on the learning process by which the rules are acquired by the rule-based model. We expand on these points next.

First and foremost, a major limitation of the function-learning literature is the absence of comparisons of rule-based versus associative-based models that focus on the learning process itself rather than on transfer performance. Advocates of the polynomial rule models never explicitly formulated a learning model, making it difficult to test these models using learning data. Comparisons of rule-based with associative-based models have been founded mainly on transfer tests after training, and no systematic comparisons have been made for these two models in terms of their ability to account for the details of the trial-by-trial learning data. It remains unknown whether or not the associative models provide a better explanation of the learning process than do the polynomial rule models. We address this gap in two ways: (1) We develop a new formal learning mechanism for rule-based models, and (2) for the first time, we test the models on the entire learning sequence during training as well as testing them in the transfer phase. This extends previous work in which transfer performance with models constrained only by the endpoint of learning has been evaluated.

Second, the question of whether or not a rule-based approach to function learning is able to capture extrapolation performance remains open. The past failures of

rule-based models to account for extrapolation, reported by DeLosh et al. (1997), were contingent on the use of a statistical learning model that assumed optimal use of past experience. Rule-based models based on less than optimal learning algorithms that are constrained to produce parsimonious representations (Busemeyer, McDaniel, & Byun, 1997; Koh & Meyer, 1991) could in principle produce imperfect extrapolation (as well as nearly perfect extrapolation for extrapolation stimuli proximal to training stimuli; Juslin et al., 2003), thereby making these rule models potentially acceptable. For instance, in casual concept learning (a more complex variant of function learning in which multiple continuous inputs predict multiple continuous outputs), Busemeyer, McDaniel, and Byun found that a parsimony mechanism was needed to account for human learning. Although advocates of rule models have proposed several ways in which parsimony might operate to constrain rule complexity (Brehmer, 1973; Koh & Meyer, 1991), these have yet to be implemented in a learning component of a rule model. Central contributions of this article are to formally implement the two parsimony principles suggested by Brehmer (1973) and by Koh and Meyer and to demonstrate the fruitfulness of these principles for improving rule models.

A third important consideration is that only polynomial rule models have been tested; other rule forms, such as the Fourier, have been suggested as alternatives (Bott & Heit, 2004) and might fare better. To provide an incisive and comprehensive test of rule-based models, in this article we develop a formal learning algorithm for rule-based models that has the potential to compete with the (associative) EXAM, and we include mechanisms that are intended to achieve parsimonious representations. We also develop models using logistic and Fourier representations, which have never been previously formalized (but see Bott & Heit, 2004, for use of a single cosine function). Finally, we extend this evaluation to broadly test both classes of models on learning and extrapolation data in a function-learning task.

THE FUNCTION LEARNING MODELS

To convincingly compare formal rule-based and associative-based accounts of function learning in light of the empirical findings, it is necessary to focus on contrasts between definitive properties and minimize incidental differences. We accomplish this by implementing both approaches within a general connectionist learning framework (cf. McClelland & Rumelhart, 1986) that employs a common background foundation. By using a common foundation, the two views can be placed on equal footing in terms of technical advantages conferred by the connectionist framework. To achieve this, both models share a set of assumptions about the cue representation, the criterion representation, and the learning algorithm used to learn the connections between inputs and outputs. The major difference lies in the use of hidden nodes to represent rules.

A core assumption of connectionist models is that all knowledge is represented by connection weights. For the rule models, conceptual knowledge about the functional relation is represented by the weights connecting input nodes to hidden nodes and the weights connecting hidden nodes to output nodes. The associative models deviate from the rule models primarily in that knowledge about the functional relation is represented solely by the weights mapping inputs directly to outputs (i.e., the hidden layer is omitted).

Within this connectionist framework, below we propose entirely new versions of rule-based models. Following that is a brief summary of the previously developed (associative) EXAM.

Rule-Based Models

The rule-based model is represented by a connectionist network that has three layers of nodes: an input layer that represents the cue, a hidden layer that represents the rules, and an output layer that represents the criterion. Each of these layers and their connections are described in turn.

Input layer. The cue value $x(t)$, presented on trial t, activates a set of input nodes for which each node, x_i, is designed to detect a possible value of the input cue. The activation of input node x_i, denoted $a_i(x)$, depends on the distance between the cue value $x(t)$ and the node x_i as follows:

$$a_i(x) = b \cdot e^{\left(\frac{x(t)-x_i}{\sigma}\right)^2}. \quad (1)$$

In this equation, σ is the standard deviation that determines the width of the generalization gradient and b is a constant used to normalize the activations (satisfying $\Sigma\, a_i = 1$). Formally, this is called a *Gaussian radial basis unit* (cf. Haykin, 1994, chap. 7), which has been used to represent input activation patterns in earlier category learning models (Knapp & Anderson, 1984; Kruschke, 1992; Nosofsky & Kruschke, 1992).

Hidden layer. Each hidden unit, denoted H_j, is interpreted as an individual component of a rule. These hidden nodes are used to compute predictions as follows. The input activation pattern flows from the input layer to the hidden layer by way of input to hidden node connections. In particular, the connection weight, w_{ji}, represents the strength of the connection from input node x_i to hidden node H_j. The prediction produced by hidden node H_j (denoted h_j) is then computed by a possibly nonlinear transformation Q of the weighted sum of input activations:

$$h_j(x) = Q\left[\Sigma\, w_{ji}\, a_i(x)\right]. \quad (2)$$

There are many types of rules that can be postulated within this approach. To provide a comprehensive evaluation of this approach, we implemented all of the known proposals suggested in the literature. Specifically, we examined what are known as *polynomial*, *Fourier*, and *logistic* types of rules. For a polynomial rule, each hidden unit computes a different trend component; for a Fourier

rule, each hidden unit computes a different cyclic component; and for a logistic rule, each hidden unit computes a different logistic response function. The key idea is that instantiation of these different rule types is achieved by the assignments of the connection weights, w_{ji} (details are provided in the Appendix).

Output layer. The hypotheses computed at the hidden layer flow to the response layer through a second layer of connections. More specifically, the hypothesized value, h_j, computed by a hidden node, is weighted by an estimate of the validity of that hypothesis, denoted v_j, and these weighted hypotheses are summed to form the prediction from the rule:

$$y(t) = \sum v_j\, h_j\,(x). \tag{3}$$

The number of hypotheses entering this sum, denoted n, determines the complexity of the model.

Conceptually, the general form of a rule is captured in this framework by the weighted combination of the hidden unit activations, which generates the prediction y. For a polynomial model, the prediction combines the trend components to form a polynomial series. For the Fourier model, the prediction combines the cyclic components to form a Fourier series. For the logistic model, the prediction combines the different logistic functions to form a logistic series.

It is well known from mathematical analysis that polynomial and Fourier series can approximate any smooth function with a sufficient number of terms. Hornik, Stinchcombe, and White (1989) have shown that the logistic series can also provide a reasonable approximation of any smooth function with a sufficient number of terms. Thus, all three types of rule models provide a sufficiently general basis for approximating the functions examined below (see Table 1A), although this does not guarantee that the appropriate weights will be learned for all training regimens.

Learning the hidden–output weights. A delta learning algorithm is used to update the validities of each hypothesis after each feedback trial, as is shown in Equation 4.

$$v_j(t + 1) = v_j(t) + \alpha\,[z(t) - y(t)]\, h_j(x). \tag{4}$$

Intuitively, this algorithm works as follows. The new validity (v) of the jth hypothesis after feedback on trial t equals the old validity plus a change. The change is the product of two parts: (1) the prediction error $e = (z - y)$ on trial t and (2) the hypothesized value h_j generated by the jth hypothesis on trial t. For example, if a hypothesis generated a large positive value for the criterion but the feedback criterion fell far below this prediction, then the validity of that hypothesis would decrease. The learning rate parameter, $\alpha > 0$, controls the amount of change in the new weight produced by feedback on each trial.

Parsimony. Recall that the delta learning algorithm was designed with one objective in mind, which is to minimize squared prediction errors. Koh and Meyer (1991) argued that the learner actually has two objectives in mind when trying to learn rules for prediction: not only to improve accuracy, but also to do so in the simplest or most parsimonious way. Parsimonious functions may generalize or extrapolate more effectively than overly complex functions.

Koh and Meyer (1991) proposed a penalty for complexity that was measured by an index of curvature—that is, a deviation from linearity. Although Koh and Meyer did not propose a learning algorithm for this penalty term, it can be implemented into the delta learning rule by making the following modification:

$$v_j(t + 1) = v_j(t) + \alpha\,[z(t) - y(t)]\, h_j(x) - \lambda_j\, v_j(t), \tag{5}$$

where the parameter λ_j represents a penalty extracted for using hypothesis H_j (i.e., a particular term of the rule, such as a cubic trend in a polynomial rule).[1]

Table 1A
Function Forms Ordered According to Learning Difficulty

Model	Function	Coefficients	MAD
Byun (1995, Experiment 1B)			
Linear	$Z = bX^{1.0}$	$b = 1.77$	.20
Square root	$Z = bX^{.50}$	$b = 1.77$	.35
Byun (1995, Experiment 1A)			
Linear	$Z = a + bX$	$a = 0.20, b = 1.77$	.15
Power, positive acceleration	$Z = a + bX^c$	$a = 0.20, b = 1.77, c = 2$	.20
Power, negative acceleration	$Z = a + bX^c$	$a = 0.20, b = 1.77, c = 0.50$	.23
Logarithmic	$Z = a + b \cdot \ln(cX + 1)$	$a = 0.20, b = 0.64, c = 15$	.30
Logistic	$Z = a + b/[1 + e^{-c\,(X - .5)}]$	$a = 0.20, b = 1.77, c = 15$	.39
Byun (1995, Experiment 2)			
Linear	$Z = a + bX$	$a = 0.15, b = 2.21$	.18
Quadratic	$Z = a - b \cdot (X - .5)^2$	$a = 1.97, b = 7.87$	.28
Cyclic	$Z = a + b \cdot \sin(cX\pi)$	$a = 1.12, b = 0.85, c = 10$	.68
DeLosh, Busemeyer, & McDaniel (1997)			
Linear	$Z = a + bX$	$a = 0.30, b = 2.21$	.10
Exponential growth	$Z = a(1 - e^{-bX})$	$a = 2, b = 4$	.15
Quadratic	$Z = a - b \cdot (X - .5)^2$	$a = 2.10, b = 8.33$	.24

Note—X ranges from 0 to 1. MAD, mean absolute deviation between the subject's prediction and the training function criterion on the first training block.

Brehmer (1973) suggested another psychological principle for introducing a parsimony principle into the learning process for rule-based models. The basic idea is to include a hypothesis in the prediction only when its validity begins to exceed a criterion in magnitude, symbolized as δ. Also, a hypothesis can be removed whenever its validity falls below this same criterion δ in magnitude.[2] This principle is closely related to the Akaike information criterion method used in statistics for model selection (see Akaike, 1973; Bozdogan, 2000), which selects a more complex model over a simpler one only when the increment in fit exceeds a criterion (two times the difference in number of parameters for the AIC index). In general, both parsimony principles may operate simultaneously to different extents, depending on the values of the parameters λ_j and δ. Note that rule learning models without parsimony are special cases in which λ_j and δ are both set to zero. In the present article, we compared models with and models without parsimony to examine the contribution of this principle.

Learning the input to hidden weights. The hypotheses for the polynomial and Fourier models are assumed to be learned on the basis of extensive past experiences in the natural world. The participants are assumed to begin the task using these previously established hypotheses about linear and quadratic trends, or slow- and fast-moving cycles. Only the validities of these hypotheses (the weights between the hidden and output layers) are learned during training on the function-learning task. Therefore, these a priori hypotheses are represented by a *fixed* set of initial weights, w_{ji}, between inputs and hidden units, which do not change during training (see the Appendix for the method used to fix these weights).

In contrast, the logistic rule model allows the hypotheses to be dynamically adjusted across trials. Accordingly, the weights w_{ji} for the logistic model are learned during training by back-propagation rather than being fixed by prior knowledge (see the Appendix for these details).

Parameters for rule-based models. Altogether, rule-based learning models entail the following model parameters: One is the standard deviation for the input nodes, σ; a second is the number of hypotheses or hidden units, n; a third is the learning rate parameter for the validities, α; and the fourth and fifth are the penalty for complexity (λ) and the cutoff threshold for including a hypothesis (δ), respectively. A sixth and seventh are used for the learning rate and momentum, respectively, but these are needed only for the logistic rule to learn the weights, w_{ji}, connecting input to hidden units. The last two parameters are not required for the polynomial and Fourier models because these weights are fixed to a priori values for these two types of rules.

Extrapolation Associative Learning Model (EXAM)

EXAM was developed to provide an extension of earlier category learning models (Knapp & Anderson, 1984; Kruschke, 1992; Nosofsky & Kruschke, 1992) to account for function learning with continuous rather than categorical responses (see Busemeyer, Byun, et al., 1997; DeLosh et al., 1997). In addition, to account for extrapolation behavior, the associative learning process was combined with a linear rule for generating extrapolations. Below, we briefly present these two components of the model.

Associative learning component. This simple component has only two layers: an input layer used to detect the stimulus cue and an output layer used to select the criterion response. The cue value presented on a particular trial produces a pattern of activation across the first layer of input nodes according to a Gaussian generalization gradient. This input activation pattern then flows through a set of linear connections directly into the second layer of response output nodes. The retrieved output activations of the response nodes are used to select the prediction for that trial. The connections from input to output are learned by a delta learning rule. (For details, see Busemeyer, Byun, et al., 1997).

It is important to note that the associative learning component implements a very general approach to function approximation called *cubic spline approximation* (Poggio & Girosi, 1990). Thus, with a sufficiently large number of input and output nodes, the associative learning component can approximate all of the functions listed in Table 1A (see Haykin, 1994, chap. 7).

Extrapolation with EXAM. The presentation of a novel extrapolation stimulus is assumed to evoke the use of a linear extrapolation response rule. The basic idea for extrapolation is illustrated in Figure 1, which can be described conceptually as follows. When presented with a novel extrapolation cue, the learner matches the novel test cue to the training stimuli, retrieves the predictions from the nearby training values, and estimates the line formed by these retrieved input–output pairs. Then, the response to the novel test cue is formed by linear extrapolation outward from the retrieved values using this estimated line. This idea is consistent with the linear extrap-

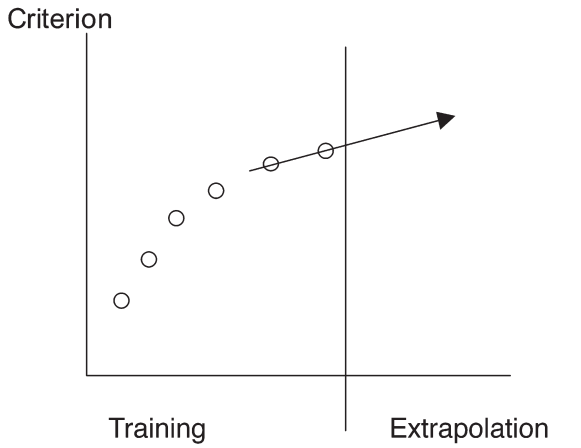

Figure 1. Illustration of the extrapolation response rule used by EXAM. The most extreme training values are retrieved from the associative network, and the slope of the extreme training values form the basis for linear extrapolation.

olation results from exponential growth curves reported by Wagenaar and Sagaria (1975). (See Busemeyer, Byun, et al., 1997, for details.)

A main difference between rule models and EXAM regarding extrapolation is the nature of the response rule for performing the extrapolation. Whereas rule-based models consistently employ a rule to generate responses to training as well as extrapolation stimuli, EXAM simply relies on the retrieved associations to generate responses to the training stimuli and evokes the linear rule only for generating responses to extrapolation test stimuli.

Parameters for EXAM. Altogether, EXAM entails the following model parameters: One is the standard deviation for the input nodes, σ; a second is the learning rate parameter for the associations, α; and the third is the standard deviation, ϕ, for generalizing extrapolation cues to training stimuli (used only during extrapolation; see Busemeyer, Byun, et al., 1997, for details).

MODEL COMPARISONS WITH BASIC FINDINGS

Study 1: Function Form and Order of Learning Difficulty

Historically, the empirical foundation for developing rule learning models for categorization was based on the order of difficulty of learning different logical relations (see Bourne, 1974). We have established a similar empirical foundation for the order of difficulty of learning different functional relations. To accomplish this, we have completed a series of experiments using a common method and procedure (see Byun, 1995; DeLosh, 1994; DeLosh et al., 1997).

The target results are presented in Table 1A. Within each experiment, the rules are listed in order of increasing difficulty, as determined by the mean absolute deviation (MAD) between the subject's prediction and the training function criterion on the first training block (see values in the last column). Accordingly, a first general challenge for the proposed models of function learning is to reproduce the order of learning difficulty for the different function forms shown in Table 1A.

Model simulations for Study 1. Simulations for these learning results were conducted for each model with the following assumption: The initial weights of the models were set to produce an initial response corresponding to the midpoint of the response scale, to reflect a neutral cover story orientation.

For the rule models, we used only the simplest versions that contained no parsimony constraints. Varying the number of hidden units did not affect the results as long as a sufficiently large number was included ($n \geq 5$ hidden units were needed for all the varieties shown in this table, but the results are shown for $n = 17$ hidden units). The standard deviation of the inputs also did not greatly affect the results as long as it was not too large (the results are based on $\sigma = 1$). The learning rate parameter did affect the results, and so this parameter was estimated separately for each model and experiment in the table.

For EXAM, we did not use the extrapolation response mechanism, and the predictions were based only on the associative learning component of the associative learning model. For this model, the standard deviation was fixed across experiments, but the learning rate was fit separately for each experiment.

The relative learning difficulty of each function form predicted by each model is shown in Table 1B. The empirically observed order (based on Table 1A) is reflected in Table 1B by the order in which the functions are listed for each experiment. Deviation between these empirical outcomes and each model's predictions is revealed for columns in which the predicted means do not increase monotonically within each experiment.[3] Examination of Table 1B shows that all of the models, with the exception of the logistic rule model, reproduced the order of difficulty reported in all experiments quite well. By contrast, for two of four experiments the results of the logistic model showed significant deviations from the results of the human learners. In Byun (1995, Experiment 1A), the positively accelerated power function was the second easiest and the logarithmic function the second most difficult of five functions for humans. The logistic model essentially reversed these, the logarithmic function being tied for second easiest and the positive power function being one of the two most difficult. Also, the logistic model violated the fundamental finding that linear functions are easier than curvilinear functions, since it learned the exponential function faster than the linear function (see DeLosh et al., 1997).

Thus, the extant empirical literature on order of learning difficulty provides a basis for disfavoring the logistic rule model. The absence of a more decisive outcome

Table 1B
Model Predictions for Order of Learning Difficulty

Model	ALM	Poly	Fourier	Logistic
Byun (1995, Experiment 1B)				
Linear	.04	.04	.05	.16
Square root	.05	.06	.06	.19
Byun (1995, Experiment 1A)				
Linear	.10	.33	.33	.17
Power, positive acceleration	.12	.37	.37	.24
Power, negative acceleration	.12	.36	.36	.19
Logarithmic	.14	.41	.41	.19
Logistic	.18	.51	.52	.33
Byun (1995, Experiment 2)				
Linear	.01	.18	.19	.12
Quadratic	.03	.31	.31	.24
Cyclic	.32	.41	.40	.68
DeLosh, Busemeyer, & McDaniel (1997)				
Linear	.04	.11	.11	.04
Exponential growth	.05	.17	.17	.02
Quadratic	.07	.27	.27	.11

Note—Each cell indicates the mean absolute deviation between the model's prediction and the training function criterion on the first trial

is probably not due to an overly limited sampling of function forms. Across four experiments, nine different function forms were tested. In the concept rule-learning literature, just four rules were used to establish a difficulty ordering that converged on a single rule-learning theory (Bourne, 1974; Shepard, Hovland, & Jenkins, 1961). By contrast, it appears that for function learning, once association and rule models are instantiated within a similar formal platform, discriminating between them will require consideration of more than one aspect of performance.

Consequently, we next examined how well the remaining models fared when applied to extrapolation data. The logistic model was discarded on the basis of its poor performance in accounting for order of function difficulty. The versions of the polynomial and Fourier models without parsimony were acceptable for order of function difficulty, and in the next section we continue our examination of these models. To foreshadow, the empirical findings for extrapolation encouraged consideration of the models with parsimony as well.

Study 2: Extrapolation

In one of the seminal studies of function learning, Carroll (1963) pointed out the importance of examining extrapolation for distinguishing associative-based versus rule-based models. Carroll reported that participants could successfully extrapolate in the direction of the training function, and he used these results to argue against associative learning models. Wagenaar and Sagaria (1975) examined extrapolation for exponentially increasing growth curves and found that although participants do extrapolate in the direction of the training function, their extrapolations seem to follow a linear rule falling short of the positively accelerated trajectory of exponential growth.

More recently, we (DeLosh et al., 1997) conducted several experiments to examine extrapolation performance. Table 2 shows results demonstrating extrapolation behavior following training on each of the last three functions shown in Table 1A. The first row of Table 2 represents the criterion for the lowest and highest training cue values for each type of function. The second row shows the criterion value generated from the training function for the extreme stimuli (used in the extrapolation tests) and for each type of training function. The third row of Table 2 represents the observed mean responses produced by the human participants for each type of extreme stimulus and type of function. As can be seen in the table, the participants extrapolated far beyond the criterion values experienced during training for all three functions (we return to the remainder of Table 2 following the next section). In sum, participants are capable of generating new responses that are outside the range of their experience and that follow the direction of trend for the training function. Because the EXAM (associative learning component), polynomial, and Fourier models could all account for the order of learning difficulty of the three function forms, we first examined the ability of these models to reproduce these basic extrapolation findings.

Model simulations for Study 2. We first tested the models that successfully captured the order of learning difficulty of Table 1A—that is, the associative learning component of EXAM (without the linear extrapolation response mechanism) and the polynomial and Fourier rule models (with $n = 17$ and without parsimony). All the models were trained using the training stimuli encountered by subjects in the DeLosh et al. (1997, Experiment 1) study. On completion of training, the models were required to perform extrapolation. To parallel the procedure for the human participants, each model was presented with 15 inputs below the lowest trained input and 15 inputs above the highest trained input, and the model produced a predicted output for each of the extrapolation inputs.

Examination of responses for the most extreme extrapolation stimuli (one at the lower end and one at the upper end) indicated that for all models the responses tended toward the midpoint of the response scale (which is the response produced by the initial weight matrices without any learning). Thus, the performances of these models diverged significantly from the human performance in DeLosh et al. (1997). The failure of the associative learning component without the use of the extrapolation response mechanism is not surprising (but see Guigon, 2004). The failure of the polynomial rule model is somewhat surprising, because this model was embraced precisely because it can support extrapolation.

Table 2
Extrapolation Responses to Extreme Transfer Stimuli

	Linear Training Function		Exponential Training Function		Quadratic Training Function	
	Lower Region	Upper Region	Lower Region	Upper Region	Lower Region	Upper Region
Training limit	1.000	1.800	1.400	1.875	1.800	1.800
Function	0.320	2.480	0.080	1.960	0.100	0.100
Observed	0.135	2.347	0.636	2.120	0.714	1.117
EXAM	0.369	2.356	0.821	1.993	1.004	1.040
Poly–pars	0.326	2.474	1.129	2.263	1.167	1.164
Fourier–pars	1.233	1.294	1.209	1.243	1.095	1.095

Note—Poly–pars, polynomial model with parsimony; Fourier–pars, Fourier model with parsimony.

The reason for the failure in this case is that we deliberately did not force the polynomial model to a simple form. Instead, the model had to learn the coefficients of the 17 trend components from experience (as humans are required to do). Many higher order trend components produce oscillation in extrapolation, with the current weightings producing responses for extreme, novel stimuli that were at the midpoint of the response scale.

If we constrain the polynomial to take on a simple form—for example, one including only linear and quadratic trends—then other problems arise. In this case, it will learn to perfectly extrapolate in the case of the linear and quadratic training functions, grossly overestimating the amount of extrapolation that humans produce. More important, the quadratic model extrapolates in the wrong direction for the exponential growth function (DeLosh et al., 1997). The lesson from this example is that simple or lower order polynomials are inadequate for capturing the wide variety of function forms that humans are capable of learning. Higher order polynomials are required for this purpose.

Previous analyses of the polynomial rule model by DeLosh et al. (1997) avoided the issue of how many trends to include in the rule in the first place. The present results show that when a rule model has to learn the number of trends (as do subjects), it does not necessarily represent that function with a minimum number of terms. By acquiring high-order forms to represent the training range, the models become generally inadequate for extrapolation. This clearly indicates the need to test the rule models with a parsimony component that adaptively learns and limits the number of terms in the models (cf. Koh & Meyer, 1991; Lewandowsky et al., 2002). Accordingly, we next evaluated the rule models with the parsimony parameters and compared their performance to that of EXAM, which includes a linear extrapolation response rule.

Simulations of EXAM and the rule-parsimony models. To give all the models the best possible opportunity to fit the extrapolation data, the polynomial and Fourier rule models (with parsimony) and EXAM were trained with parameters that were estimated to maximize the fits directly to the transfer test data of DeLosh et al. (1997). The rule models were fit using $n \geq 3$ hidden nodes and four parameters (the standard deviation of the generalization gradient, a learning rate parameter, and two parsimony parameters). Note that the parsimony mechanisms could reduce the number of hidden nodes that were operative. EXAM was fit using three parameters (the standard deviation of the inputs, the learning rate parameter, and the standard deviation for generalizing to extrapolation stimuli).

The rows labeled "EXAM," "Poly–pars," and "Fourier–pars" in Table 2 provide the values of the models' outputs for the most extreme lower and upper extrapolation stimuli. As can be seen in the table, an important finding from this analysis is that the Fourier rule model still does not extrapolate beyond trained criterion values for the linear function and for the upper extrapolation region of the exponential function (compare the row labeled "Fourier–pars" with the first row in Table 2). These results held for a wide variety of numbers of hidden units ($n = 3$ to $n = 17$). These results disfavor the Fourier rule model as a fruitful account of human function learning, since human learners extrapolate beyond the trained values in directions appropriate for the training functions.

By contrast, the polynomial rule model with parsimony extrapolated beyond the training criterion values and did so in the direction of the function for all of the function forms (compare the row labeled "Poly–pars" with the first and second rows in Table 2). The results shown in Table 2 are based on setting the number of hidden nodes to 17, but similar results were obtained with a smaller number of hidden nodes (e.g., $n = 5$). In an even more telling result, the pattern of the extrapolation from the polynomial rule model dovetails fairly well with the subject performances shown in the table (third row).

The result described above has significant theoretical implications. DeLosh et al. (1997) constrained the polynomial rule model to the lowest power (extreme parsimony) needed to achieve the level of accuracy produced by participants at the end of training. Essentially, this produced polynomial rules that were biased a priori to match the polynomial expression of the functions being learned. Accordingly, by the end of training for the linear and quadratic functions, the polynomial expressions of the rule model perfectly paralleled these target functions. The consequence was that in extrapolation, the polynomial rule model extrapolated perfectly along the target function, thereby diverging from human performance. On the basis of this divergence, DeLosh et al. concluded that rule models cannot adequately account for human extrapolation in function learning. As we anticipated in the introduction, the present modeling results demonstrate that the rejection of rule models on these grounds was premature. A more general polynomial rule model formalized to incorporate parsimony produced imperfect extrapolation that better approximated the extrapolation produced by participants in function-learning experiments. The value of including parsimony factors in rule learning models of function learning is also paralleled by similar modeling efforts in intervening concept learning (Busemeyer, McDaniel, & Byun, 1997).

The last important result was that the associative-based EXAM (see the row labeled "EXAM" in Table 2) also extrapolated at both ends of the three function forms, in the appropriate direction and with a topography that overlapped nicely with that seen for the human subjects. Thus, both the associative-based EXAM and the rule-based polynomial model captured critical aspects of human extrapolation performance required for a model of function learning. To attempt to further distinguish between these two remaining models, our final study was an analysis of the models' ability to account for the pattern of subjects' predictions throughout learning and subsequently to account for the pattern of subjects' extrapolation responses when the models were transferred to the extrapolation trials.

Study 3: Pattern of Responses During Function Learning and Extrapolation

Thus far in the function-learning literature, no detailed model comparison has been conducted on the basis of the learning patterns evidenced during the entire learning sequence. Thus, it remains uncertain whether the polynomial–parsimony rule model or the associative model (EXAM) can account for the trial-to-trial function-learning patterns observed for human learners.

To conduct these model tests, we applied the polynomial rule model (with parsimony) and EXAM (with linear extrapolation) to detailed learning and extrapolation data reported by DeLosh (1994) for a negative linear function and a quadratic function. In this study, all of the participants were presented with exactly the same training sequence: Within each block of 10 trials, the same systematically increasing sequence of cue values was presented to each person.

Four parameters of the polynomial rule model were used (standard deviation of the inputs, learning rate, and two parsimony coefficients) to fit the learning data. To provide a comprehensive examination of the polynomial rule model, we fit models with hidden units ranging in number from 2 to 17. For EXAM, only two parameters were used (standard deviation of the inputs and the learning rate; for extrapolation, we simply set the standard deviation for generalization equal to that for training stimuli). For both models, the parameters were estimated from the learning data by minimizing the sum of squared error. Model performance was measured by $R^2 = 1 - SSE/TSS$, where SSE = sum of squared deviations from model predictions and TSS = total sum of squared deviations around the mean. These parameters were then used to generate predictions for the transfer test. Thus, importantly, a priori predictions of each model were used to test interpolation and extrapolation performance. The results for each function are reported in turn.

Negative linear condition. The learning data for the negative linear condition are shown as the open circles in Figure 2 (and again in Figure 3), where each panel presents performance on 10 consecutive trials. The best-fitting polynomial model was obtained with $n = 7$ ($R^2 = .96$, $\alpha = 34.71$, $\sigma = .21$, $\lambda = .3125$, $\delta = 0$), and the asterisks in Figure 2 indicate the predictions for this model. The asterisks in Figure 3 show the predictions for the learning performance of EXAM ($R^2 = .94$, $\alpha = .3056$, $\sigma = .0553$). As can be seen by a comparison of Figures 2 and 3, both models did a good job of accounting for learning of the negative linear function.

Quadratic condition. The learning data for the quadratic condition are shown as the open circles in Figure 4 (and also in Figure 5). Once again, the best-fitting poly-

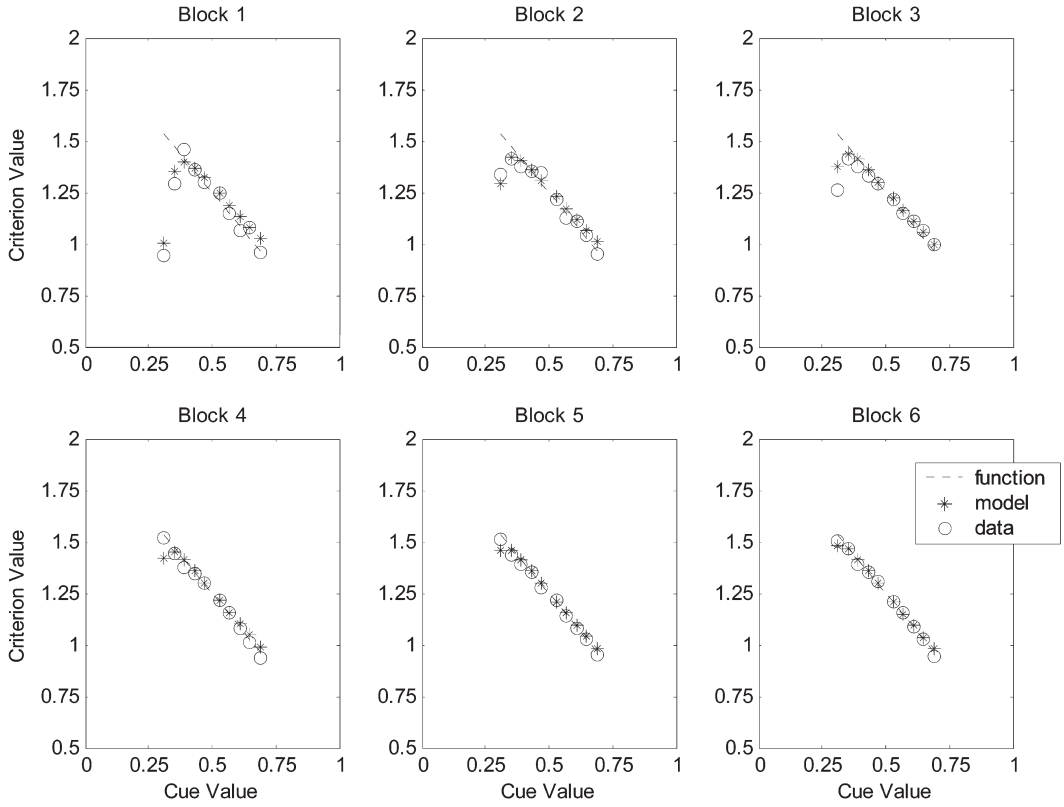

Figure 2. Learning performance on the negative linear function for the best-fitting polynomial model with parsimony (hidden nodes set to $n = 7$) and corresponding data from the learning phase of DeLosh (1994). The data are aggregated in six blocks with 10 trials per block.

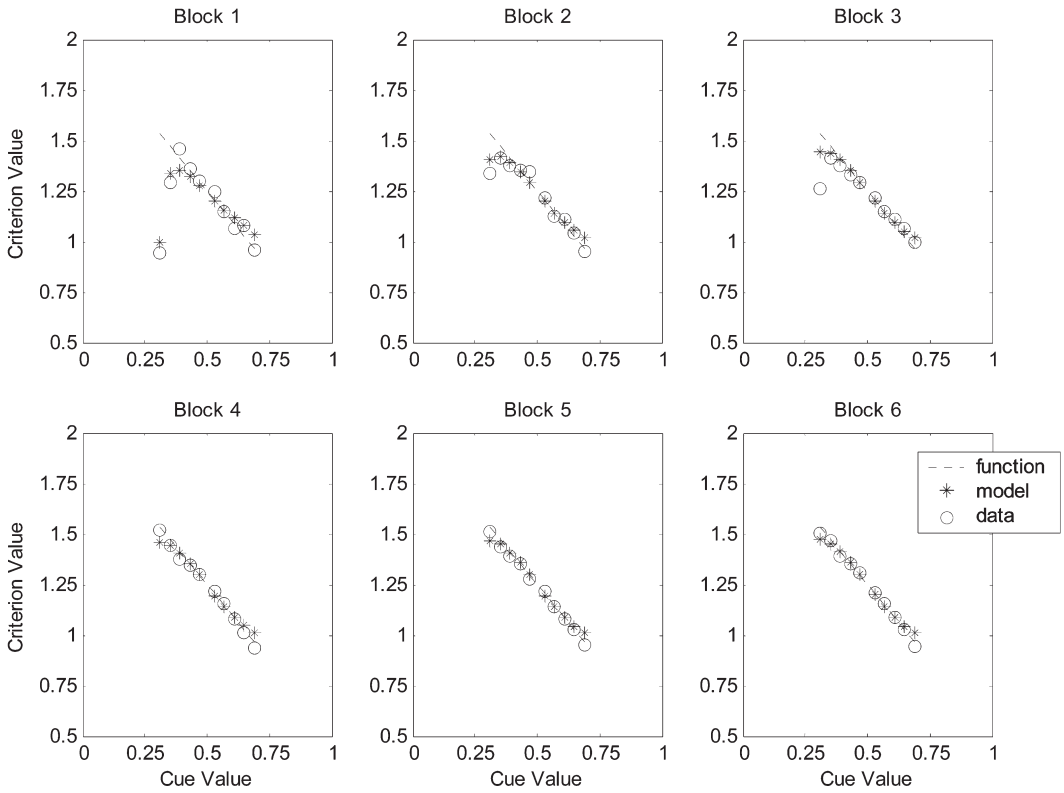

Figure 3. Learning performance on the negative linear function for EXAM and corresponding data from the learning phase of DeLosh (1994).

nomial rule model was obtained with $n = 7$ ($R^2 = .70$, $\alpha = 17.27$, $\sigma = .01$, $\lambda = .3372$, $\delta = 0$). It is interesting to note that the quadratic model (with $n = 3$) fit much worse ($R^2 = .60$, $\alpha = 19.44$, $\sigma = 1$, $\lambda = .0113$, $\delta = 0$). The predictions of the polynomial rule model (with $n = 7$) are shown as the asterisks in Figure 4. As can be seen, the polynomial rule model did not learn as quickly as the humans did during the first block of training. The predictions from EXAM are shown as the asterisks in Figure 5. By contrast, EXAM ($R^2 = .81$, $\alpha = .5387$, $\sigma = .051$) more closely approximated the learning pattern of the participants (see Figure 5), although not perfectly.

Transfer performance. Subsequent to learning, these models were transferred to interpolation and extrapolation trials. Figures 6 and 7 provide the participants' performances (open circles) and model predictions (asterisks). Considering first the polynomial rule model, Figure 6 confirms the suggestion above that once the rule model is extended with a learning component and is fit to the trial-by-trial learning data, this model will not necessarily extrapolate perfectly along the trained function. For the linear function, the polynomial rule model produces a fairly good fit to participants' linear extrapolation ($R^2 = .92$). However, the model predictions begin to tail off from the linear function in a curvilinear (cubic) fashion, whereas the participants' extrapolation appears to be more linear. For the quadratic function, surprisingly, the polynomial rule model showed very little extrapolation as opposed to human learners ($R^2 = -.54$)[4] and very little evidence of acquisition of a quadratic function.

Next, consider EXAM, which assumes a linear extrapolation response from nearby training associations, as shown in Figure 7. For the linear training function, EXAM captures the lower extrapolation region fairly well. On the upper extrapolation region, however, EXAM extrapolated above the assigned function, but participants extrapolated below the assigned function. The overall fit for EXAM ($R^2 = .89$) was only slightly below that obtained for the polynomial rule model for the linear function. The pattern of fits changed when the quadratic functions were considered. In this case, EXAM provided a much better account for extrapolation ($R^2 = .53$), in that participants again seemed to show fairly linear extrapolation. As can be seen in Figure 7, EXAM's linear extrapolation was not as steep as that produced by participants or the assigned function. The shallow linear extrapolation displayed by EXAM probably reflects inadequate learning of the extreme input–output points given in training (see DeLosh et al., 1997).

Accordingly, we trained both EXAM and the polynomial rule model (with $n = 7$ hidden units) until they produced highly accurate predictions for the training criterion values and then transferred them to the extrapolation stimuli. These results are shown in Figures 8 and 9 for

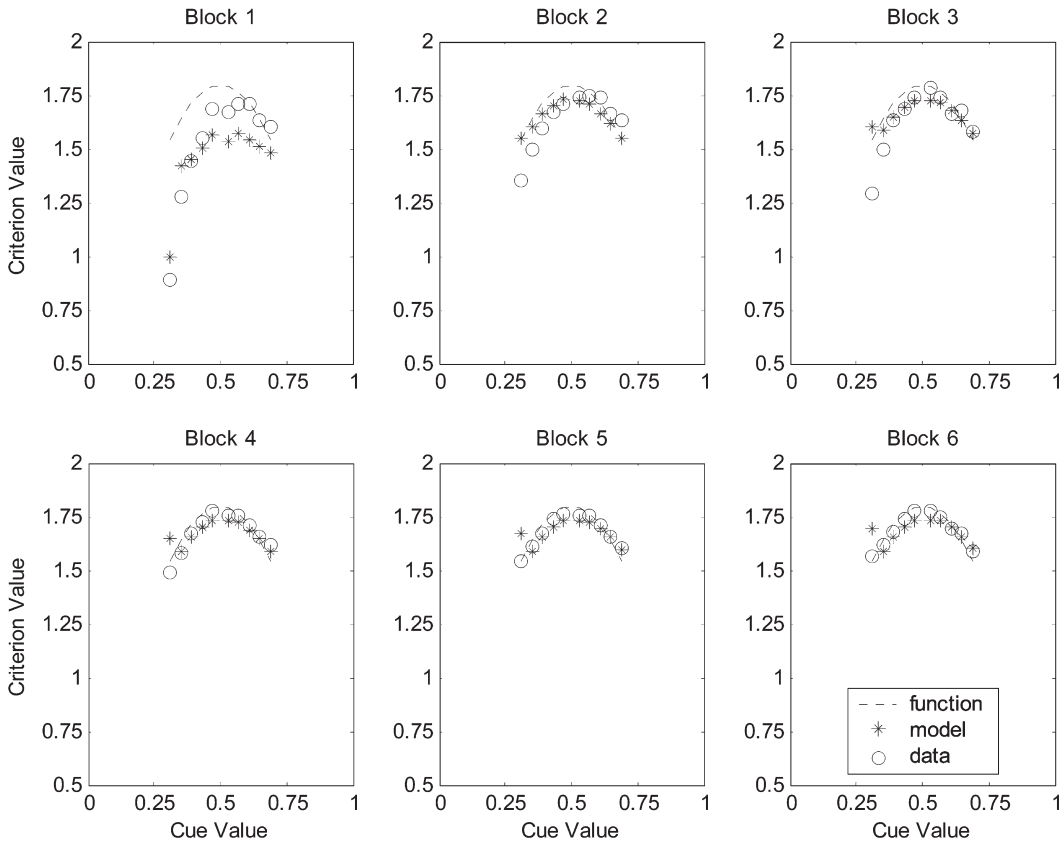

Figure 4. Learning performance on the quadratic function for the best-fitting polynomial model with parsimony (hidden nodes set to $n = 7$) and corresponding data from the learning phase of DeLosh (1994).

the polynomial rule model and EXAM, respectively. As can be seen in Figure 8, when the polynomial rule model is trained to produce highly accurate predictions for the training criterion values, it again fares well for the linear function ($R^2 = .95$). However, for the quadratic function, although its predictions are much improved relative to those above, the rule model does not completely capture the human data ($R^2 = .87$). The rule model's divergence is seen as the extrapolation points get farther from trained values, for in these cases the model produces an inappropriate curvilinear upswing. By contrast, as can be seen in Figure 9, EXAM's quadratic extrapolation better resembles the human data ($R^2 = .92$), and its linear extrapolation almost perfectly captures the human data ($R^2 = .96$) as well.

GENERAL DISCUSSION

Recent work has challenged the long-standing assumption that rule acquisition underlies the conceptual basis of human function learning (Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991). DeLosh et al. (1997) found that a popular polynomial rule model of function learning did not produce extrapolation, as did human learners. Because the DeLosh et al. study provided preliminary evidence (in terms of extrapolation performance) countering a particular polynomial rule model, some theorists have been tempted to discard polynomial rule models altogether (cf. Kalish et al., 2004; Lewandowsky et al., 2002). The dismissal of rule models on the basis of a single study of one particular rule model is premature, however (see, e.g., Juslin et al., 2003). There are a number of possible rule forms, most of which have not been formalized or evaluated. Moreover, no rule model has yet implemented a psychologically based learning component, and, accordingly, the instantiation of any particular rule examined in models has been constrained by theorists rather than by model-specified learning. We comprehensively addressed these issues by developing several new rule models that incorporated learning mechanisms with parsimony, and we considered two additional rule forms (Fourier and logistic) not examined in previous models. We first discuss the results of the modeling and evaluation of rule models, and then those of EXAM.

Rule Learning

Our results clearly established that the form of the rule is critical in terms of its being able to account for human function learning. The logistic rule that we examined

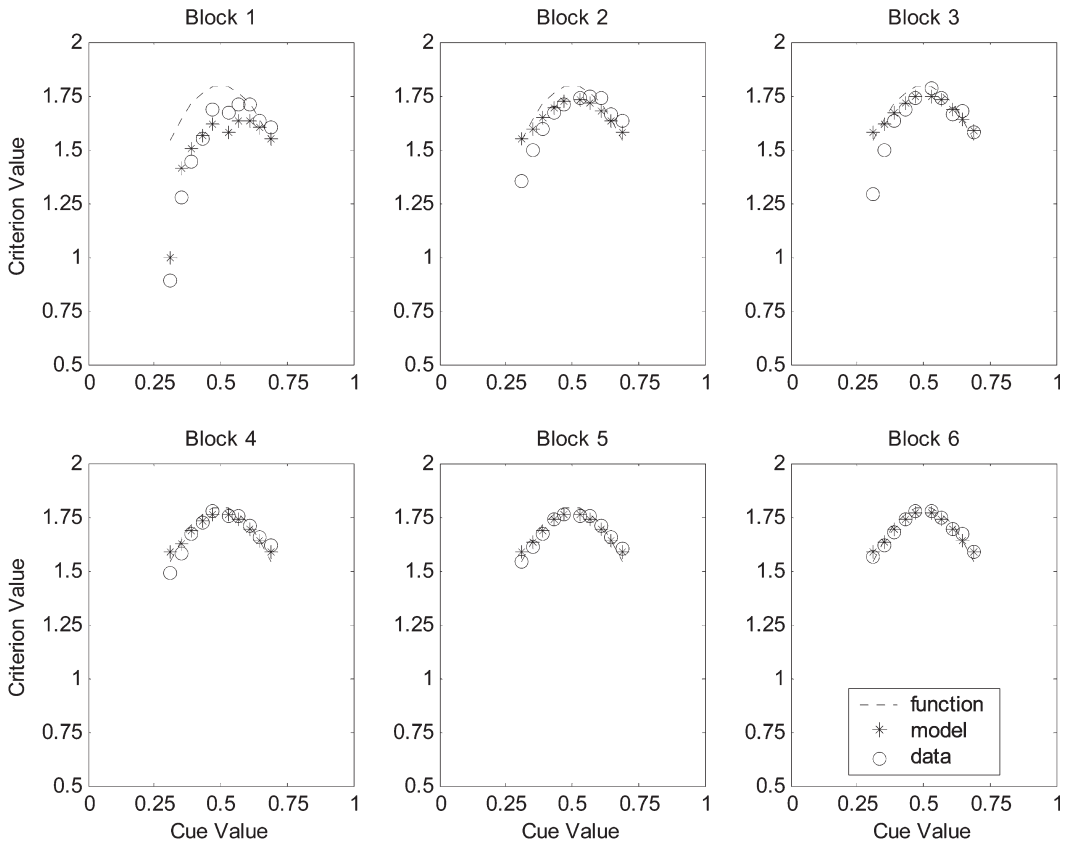

Figure 5. Learning performance on the quadratic function for EXAM and corresponding data from the learning phase of DeLosh (1994).

could not handle benchmark results of order of learning difficulty across function forms. The Fourier and polynomial rule-learning models, however, were able to account for order of learning difficulty (Study 1).

Because the learning difficulty data could not completely legislate among all of the candidate rule-learning models, we next considered extrapolation (transfer) performance. None of the rule models without parsimony could account for the basic patterns of human extrapolation behavior (see Study 2). Extreme parsimony as well produces important divergences from human function-learning patterns. DeLosh et al. (1997) imposed the constraint that the rule models adopt the lowest possible number of terms in learning linear and quadratic functions. With this constraint, the rule model was limited to a linear form for the linear function and a quadratic form for the quadratic function, and subsequently produced extrapolation that exactly reproduced the intended function. In a similar vein, Juslin et al. (2003) found that a minimal-term rule abstraction model produced extrapolation that followed the trained function and, importantly, that this model accounted for human extrapolation performance in a multiple-cue-learning task with continuous responses. In Juslin et al., extrapolation testing was limited to two examples, one near the lower ends and one near the upper ends of the trained examples. In contrast, DeLosh et al. tested extrapolation across a range of untrained values. Here, human learners' extrapolation increasingly deviated from the intended function as the extrapolation stimuli became more distant from trained points, a result that has been obtained in other studies in which extrapolation was examined across a range of functions (Byun, 1995; DeLosh, 1994). Thus, accurate extrapolation for new stimuli proximal to trained points may not prove decisive in evaluating the merits of various rule models.

Critically, the present work indicated that when parsimony was included as one of the objectives during learning (but not the overriding component) the polynomial rule model (and the Fourier model) did not extrapolate perfectly, in line with human extrapolation behavior. Thus, perfect extrapolation is not necessarily inherent to rule models, so the absence of perfect extrapolation cannot be used to dismiss rule-learning models of function learning in general. The importance of including parsimony in the learning component of rule models supports and converges with formal theoretical work in another complex concept-learning domain, that of intervening concept learning (in which multiple continuous inputs are associated with multiple continuous outputs; Busemeyer, McDaniel, & Byun, 1997). Our formal instantiation of

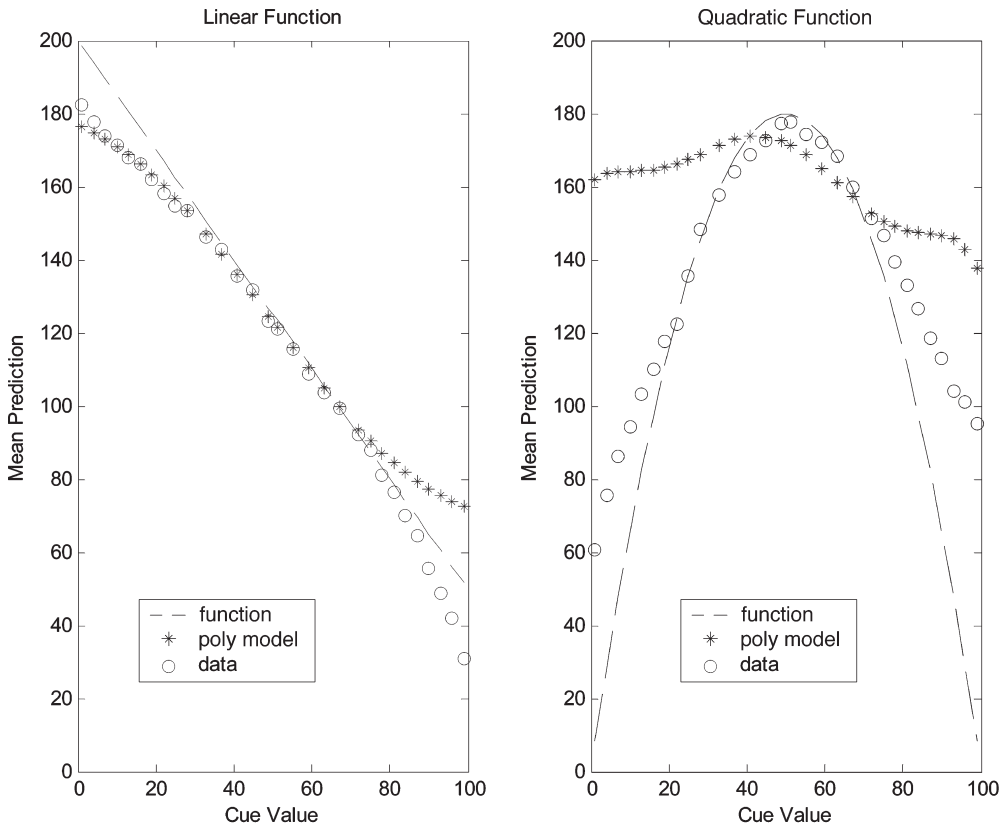

Figure 6. Transfer performance (interpolation and extrapolation) on the negative linear (left panel) and quadratic (right panel) functions for the best-fitting polynomial rule model with parsimony (hidden nodes set to $n = 7$), along with the interpolation and extrapolation data from the transfer phase of DeLosh (1994). The extrapolation responses are those with cue values below 30 and above 70.

parsimony also appears to dovetail with Lewandowsky et al.'s (2002) suggestion that expediency is an important principle of human function learning. Expediency is the desire to find an efficient solution, and rules with fewer terms (higher parsimony) represent simpler, more efficient approximations of an unknown function.

The new rule models with parsimony still produced some divergence from human extrapolation, however. In Study 3, for both the linear and the quadratic functions, for extrapolation points that were distant from trained points the polynomial rule model (the only rule model not already disfavored by Studies 1 and 2) began to produce predictions that curled back toward the trained output values of the intended function. In contrast, human learners continued to produce more extreme outputs, and did so in a more linear fashion.

For the first time, we also examined learning patterns for the rule models in detail. In some cases, a rule model may fare well in accommodating human performance when fits are restricted to asymptotic learning performance and the learning pattern is ignored (cf. Juslin et al., 2003). We considered the best polynomial rule model over the entire learning process, and this model was not very successful in accounting for human learning performance for a quadratic function in either a quantitative ($R^2 = .70$ vs. $R^2 = .81$ for EXAM) or a qualitative aspect. Human responses seemed to approximate somewhat a quadratic shape in the first 10 trials, with some responses overlapping with the criterion (see Figure 4). At least 20 more trials were needed to adjust these predictions to a more symmetrical quadratic form. By contrast, the polynomial rule model was accurate in approximating only few, if any, of the intended criterion values in the first 10 learning trials (see Figure 4), but then quickly acquired a symmetrical quadratic form in the next 10 trials.

In sum, the best rule model of the five examined—namely, the polynomial rule model with parsimony—although not exhibiting the problem reported for previously formalized rule models, did fall short of adequately reproducing human function learning and extrapolation performance on other dimensions. Unless other types of formal rule models can be successfully implemented, it appears that a more fruitful candidate for a model of function learning is needed.

EXAM

Busemeyer, Byun, et al. (1997) proposed a hybrid model of function learning (EXAM) in which learning of the experienced input–output points was accomplished

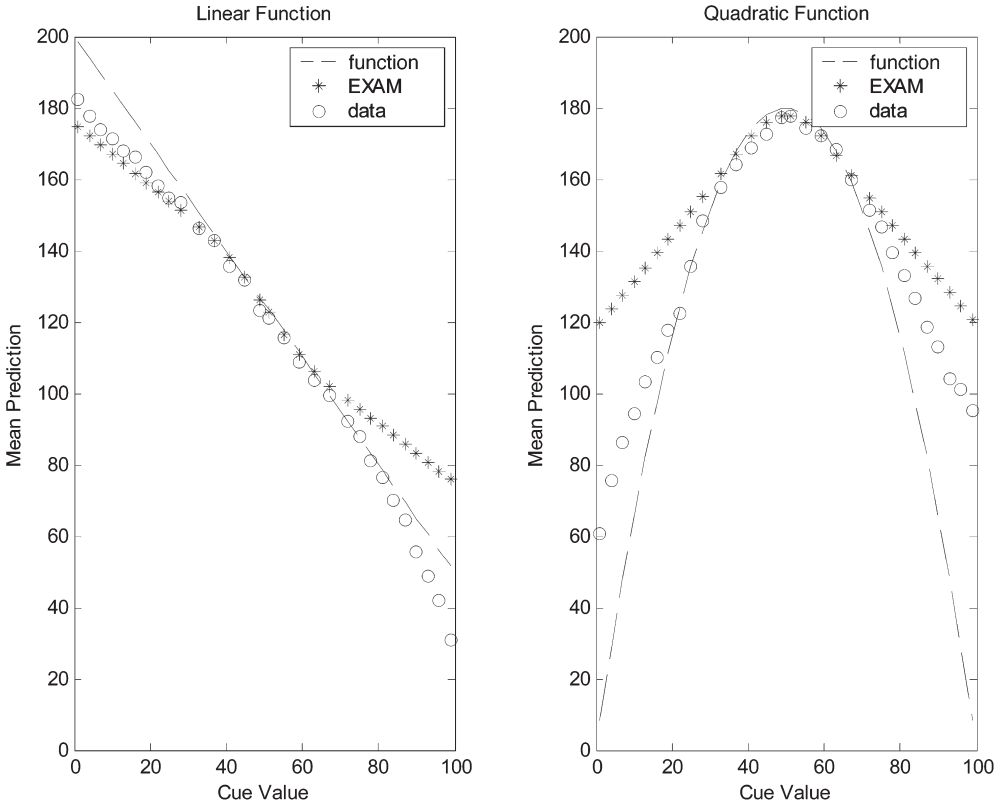

Figure 7. Transfer performance (interpolation and extrapolation) on the negative linear (left panel) and quadratic (right panel) functions for EXAM and the corresponding transfer data from DeLosh (1994). The extrapolation responses are those with cue values below 30 and above 70.

with associative mechanisms similar to those proposed in earlier category models (Knapp & Anderson, 1984; Kruschke, 1992; Nosofsky & Kruschke, 1992) and extrapolation was accomplished with a linear response generation mechanism (Wagenaar & Sagaria, 1975). An initial study in which EXAM was tested with regard to human extrapolation behavior showed that humans exhibited linear extrapolation (after learning training points presented with a neutral cover story) and that EXAM approximated this extrapolation behavior well (DeLosh et al., 1997). Although their results were encouraging, DeLosh et al. did not evaluate EXAM's learning performance or extrapolation performance after simulation of the entire learning session. The present study provided a more extensive test of EXAM through examination of these dimensions of human function learning.

We were able to show that EXAM consistently accounts for patterns of acquisition in human function learning and that it does so better than rule models instantiated in a similar formal platform. Study 1 showed that EXAM accounted for the order of difficulty in learning different function forms. In a challenging competitive test of EXAM versus the polynomial rule model with parsimony, Study 3 applied the models to detailed trial-by-trial learning data for a negative linear function and a quadratic function. For the negative linear function, EXAM almost perfectly captured the human learning data, but so did the polynomial rule model. EXAM did not approximate the human learning data for the quadratic function quite as well, but it did so better than did the polynomial rule model.

As an extension of this competitive test, after both models were fit to trial-by-trial learning data, the models were transferred to extrapolation without adjustment of parameters, thereby allowing genuine competitive predictions (cf. DeLosh et al., 1997; Juslin et al., 2003). For the quadratic function, EXAM accounted for a significantly larger proportion of the variance in extrapolation than did the polynomial rule model (+.44 for EXAM vs. −.54 for the polynomial rule model when only the extrapolation points were considered). Furthermore, the linear extrapolation produced by EXAM appeared to capture the topography of human extrapolation better than did that produced by the rule model. In agreement with this claim, when the models were trained until they achieved maximum learning accuracy, EXAM but not the rule model almost perfectly reproduced human extrapolation performance for the quadratic function.

Thus, the present study supports EXAM as the most viable formal model of function learning thus far presented in the literature. Still, as we have just noted, EXAM did not comfortably span the learning and extrapolation data for the quadratic function with regard to being able

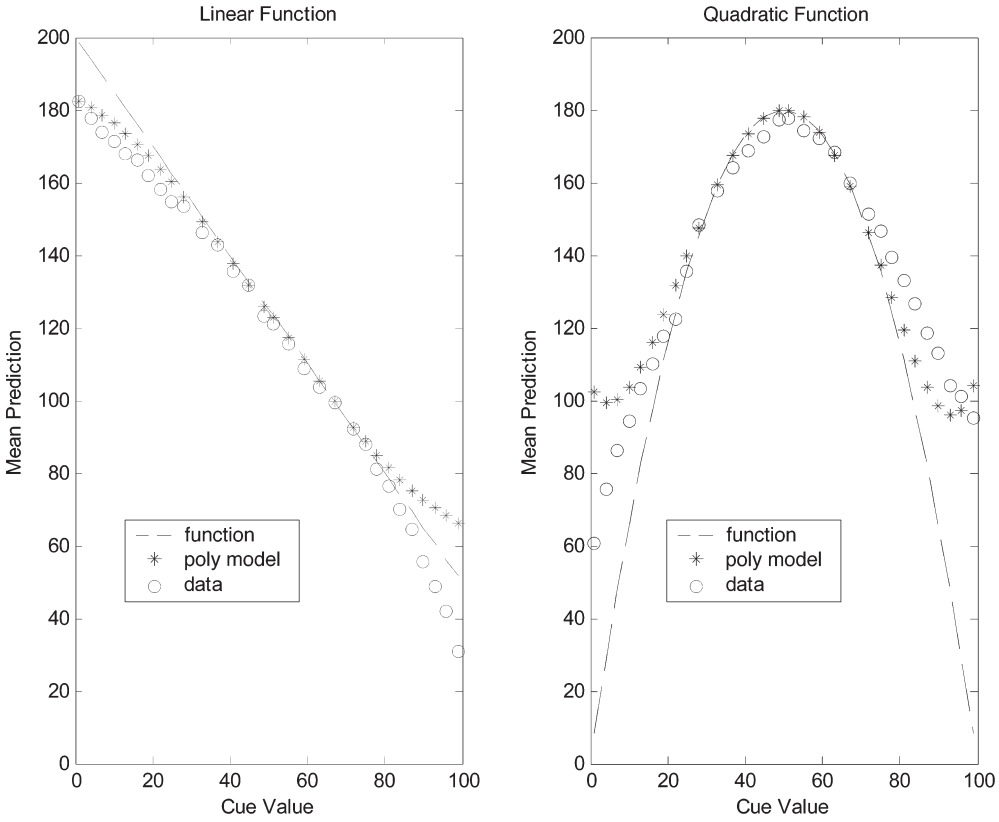

Figure 8. Transfer performance (interpolation and extrapolation) on the negative linear (left panel) and quadratic (right panel) functions for the polynomial rule model after accuracy at end of learning is maximal (with $n = 7$ hidden units), along with the transfer data from DeLosh (1994).

to transfer from the simulated learning state to a good quantitative approximation of human performance (see Study 3), thereby suggesting that some modifications might be fruitful. Perhaps some fine-tuning of starting values or learning parameters (e.g., addition of a parsimony parameter) is needed for EXAM to better account for extrapolation after simulating the human learning performances. Or, perhaps the assumption that the linear output response mechanism is based on generalization to nearby training points needs some modification. Clearly, none of these changes would alter the basic theoretical approach embodied by EXAM.

New Developments

Very recently, a number of new models for function learning have appeared (Bott & Heit, 2004; Guigon, 2004; Kalish et al., 2004).[5] We briefly evaluate the strengths and weaknesses of these new developments in comparison with EXAM.

Mixed associative and rule model. On the basis of their experimental findings that participants produced nonmonotonic extrapolation for a learned cosine (cyclic) function, Bott and Heit (2004) argued that EXAM was incomplete. In its stead, they proposed a hybrid model that consists of an exemplar module similar to the associative learning component of EXAM, plus a rule module based on a cosine function. This cosine-based rule module supported the nonmonotonic extrapolation produced by the participants. An advantage of this model is that it is flexible enough to account both for the linear extrapolation observed in previous studies and for the nonmonotonic extrapolation performance observed by Bott and Heit. However, it remains to be seen whether a dual module approach captures the psychological processes underlying the behavior reported by Bott and Heit.

First, as acknowledged by Bott and Heit (2004), their model does not yet include a learning algorithm. Second, our original EXAM may capture the nonmonotonic prediction behavior outside the training range if we assume that once participants have learned the periodic nature of a mapping, they recode the inputs at the end of each period to repeat the cycle. For example, when learning to predict weather across time, humans could learn to recode time into months that range from 1 through 12 after each period of a year. Thus, predicting the weather in the first month of 2003 yields virtually the same input as predicting the weather in the first month of 2004. These two time points are treated as the same inputs rather than as an extrapolation into a new temporal region. Essentially, this process would not require nonmonotonic extrapolation.

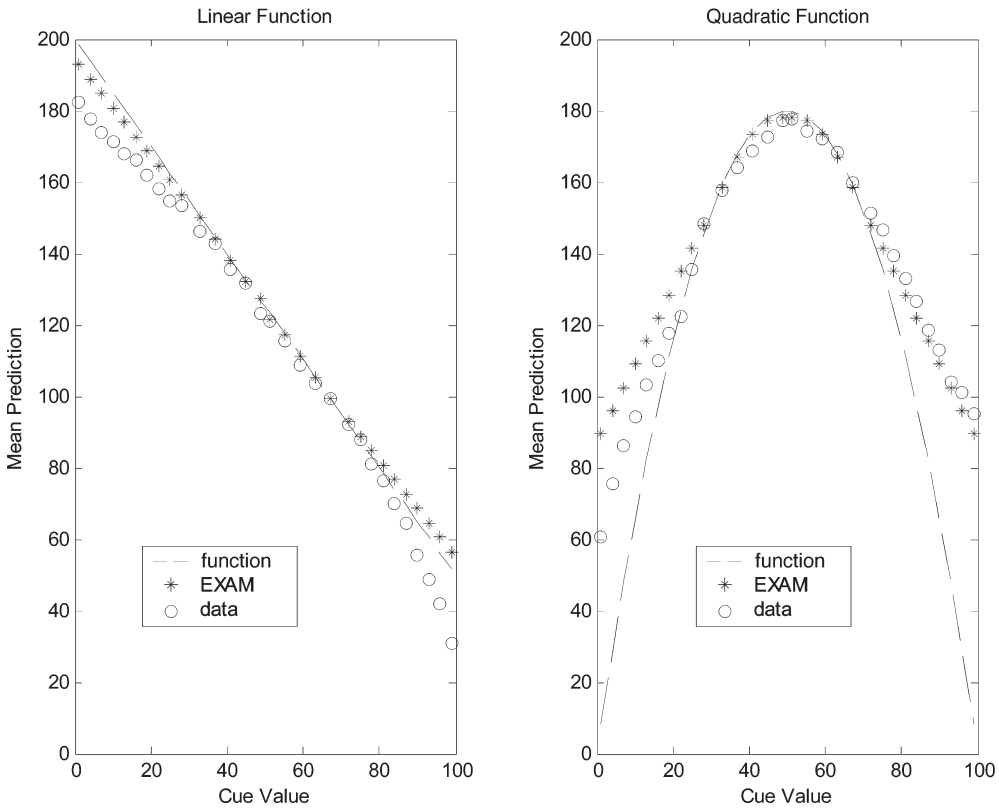

Figure 9. Transfer performance (interpolation and extrapolation) on the negative linear (left panel) and quadratic (right panel) functions for EXAM after accuracy at end of learning is maximal, along with the transfer data from DeLosh (1994).

Neural network model. Guigon (2004) recently proposed a neural network model of function learning that postulates an associative learning approach similar to that used in EXAM, except that a cumulative input activation function is used rather than the Gaussian radial basis function employed by EXAM (see Guigon, 2004, for details). Guigon has shown that this model can accommodate the basic findings reported by Busemeyer, Byun, et al. (1997), as well as some additional perceptual motor learning findings. The advantage of this model is that it does not require an additional rule-based extrapolation mechanism as postulated in EXAM; instead, it can extrapolate beyond experience simply by using the associative network. The disadvantage of this approach, however, is that training the network is very slow, so that the model cannot learn within the same number of training cycles as humans can. Note that the associative learning model used in EXAM provides a very good approximation to the learning rates of humans, at least for the functions examined in Study 3 of the present article.

Function partitioning. Lewandowsky et al. (2002) proposed a nonformalized theory of human function learning that relies on partitioning the function into different segments and "either finding the one [linear] function that works for all stimuli or putting together a piecewise-linear approximation of the correct function" (p. 191). This view was motivated by experiments in which various segments of the function were associated with different contextual cues, with these cues appearing to promote segmentation of the function in a manner consistent with the proposed partitioning approach. Lewandowsky et al. argued that an extension to EXAM based on current associationistic category models (ALCOVE) would not be able to accommodate their results showing participants' reliance on contextual cues to mediate function learning. Modifications of EXAM's assumptions regarding interpolation and extrapolation could possibly accommodate those results. This challenge is left for future research.

EXAM seems to straightforwardly accommodate Lewandowsky et al.'s (2002, pp. 190–191) core assumption that the desire to acquire an efficient solution to a problem—that is, expediency—is pervasive in function learning (which is the claimed motivation for knowledge partitioning). First, with regard to Lewandowsky et al.'s particular findings, the suggestion is that participants associated a particular context (cover story) with a particular kind of function that was then applied in that context. In EXAM, the knowledge associated with particular cover stories is implemented by setting the initial weights $W(0)$ equal to the weights obtained from prior training on a function consistent with the cover story. For example, if participants are told to predict the height of a ball

as a function of the time it is in the air, then the initial weights are obtained from preliminary training on inverted U-shaped quadratic functions in consistency with the cover story. A second assumption appears to be that expediency may bias subjects toward linear and positive functions. If such were the case, then initial weights in EXAM could easily be set to produce a linear and positive bias rather than the neutral weights used in the present and previous works (DeLosh et al., 1997) with neutral contexts (cover stories).

The partitioning approach was amplified and formalized by Kalish et al. (2004) in a new model of function learning, the *population of linear experts* (POLE) model.[6] POLE specifies how stimuli and responses are partitioned into independent linear mappings under conditions like those examined in most existing function-learning paradigms, in which the context does not provide clear cues for partitioning the function into segments. Briefly, when confronted with a function-learning task, POLE activates a large number of linear functions ($N = 64$ in Kalish et al., 2004) that provide the basis for producing responses during learning and underlie the acquired knowledge about the target function. When a cue value is presented on a trial, each of the linear functions produces a prediction. The selection of the prediction that serves as the response on the trial is probabilistically determined on the basis of the acquired strength of association between the cue value and each function and on that of the strength of the association between the context cues (if context cues are present, as in Lewandowsky et al., 2002) and each function.

A critical implication of this model is that across trials responses to a cue value will be multimodally distributed. The idea is that on different trials the response can be produced by a different linear function, and so responses should cluster around these different competing linear functions. Kalish et al. (2004) tested this prediction in two experiments using functions based on conflicting separated linear segments (e.g., vertically offset separated functions). In consistency with POLE but not with EXAM, Kalish et al. found that when learners were transferred to untested cue values that were between the values of the endpoints of the trained linear segments, responses were multimodally distributed. Kalish et al. also found that POLE fared well when evaluated against benchmark function-learning findings regarding function difficulty (see Busemeyer, Byun, et al., 1997; Study 1 of the present article). Finally, POLE accounted for the transfer (extrapolation) findings of DeLosh et al. (1997) focused on in the present Study 2.

Accordingly, support for POLE is thus far impressive. We believe there are notable limits as well. In the experiments reported by Kalish et al. (2004), training consisted of only three trials (across blocks) of over 50 cue values (Experiment 1). In many function-learning situations, including real-world examples identified by Kalish et al. (e.g., how much to water the lawn as a function of temperature), it is unlikely that the learner encounters or encodes over 50 cue values for the function. Arguably, in everyday function learning the learner more characteristically encounters (or encodes) a handful of cue values and perhaps encounters these values numerous times (cf. the paradigm of DeLosh et al., 1997). Furthermore, in Kalish et al.'s paradigm, participants are informed that their response is accurate if it is within four units of the function response; such feedback is less fine-grained in this paradigm than in others. Under any or all of these alternative conditions, it is possible that learning processes assumed by EXAM prevail. Moreover, the behavior directly established by Kalish et al.'s results is that transfer to untested values between endpoints of conflicting linear segments produces uncertainty about which segment to activate to support transfer. In these special cases, EXAM might be modified to display uncertainty in terms of which learned values are applied to the extrapolation response rule.

A more significant limitation is that, to account for the extrapolation performances reported in DeLosh et al. (1997), POLE's parameters were fit directly to the transfer data. POLE's representation is based on linear functions, so it is not surprising that POLE could fit the linear extrapolation observed by DeLosh et al. POLE has not yet been shown to predict extrapolation performance, as does EXAM. That is, it is uncertain how POLE would succeed relative to EXAM if the model were additionally challenged to first learn the training stimuli (as in Study 2) or to simulate detailed learning performance (as in Study 3) before attempting to account for extrapolation. Furthermore, it is uncertain that POLE would even produce learning topographies reported by others (e.g., Byun, 1995; DeLosh, 1994) and modeled in the present study.

Conclusions

The significance of the present study is that it represents the most comprehensive specification and evaluation to date of formal models of function learning, with the results arguing strongly against long-standing rule models. An alternative formal hybrid model, EXAM, appears promising. The theoretical appeal of EXAM is that, by assuming a basic associative learning mechanism, it integrates theories of human categorization learning and function learning to provide a general approach to a range of human conceptual behaviors. The study also reinforces the fruitfulness of formalizing models and conducting comprehensive tests of models using benchmark data, since such formalization and testing provide clear markers for the capabilities and shortcomings of function-learning theories.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Akadémiai Kiadó.

Ashby, F. G., & Maddox, W. T. (1998). Stimulus categorization. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 251-301). New York: Academic Press.

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 38-50.

Bourne, L. E., Jr. (1974). An inference model of conceptual rule learning. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 231-256). Potomac, MD: Erlbaum.

Bourne, L. E., Jr., Ekstrand, B. R., & Dominowski, R. L. (1971). *The psychology of thinking*. Englewood Cliffs, NJ: Prentice-Hall.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62-91.

Brehmer, B. (1973). Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior & Human Performance*, **9**, 377-395.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior & Human Performance*, **11**, 1-27.

Busemeyer, J. R., Byun, E., DeLosh, E., & McDaniel, M. A. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 405-435). Hove, U.K.: Psychology Press.

Busemeyer, J. R., McDaniel, M. A., & Byun, E. (1997). The abstraction of intervening variables from experience with multiple-input multiple-output causal environments. *Cognitive Psychology*, **32**, 1-48.

Byun, E. (1995). *Interaction between prior knowledge and type of nonlinear relationship on function learning*. Unpublished doctoral dissertation, Purdue University.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua*. Princeton, NJ: Educational Testing Service.

DeLosh, E. L. (1994). *Rule abstraction and hypothesis testing in the learning of functional concepts*. Unpublished master's thesis, Purdue University.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 968-986.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Guigon, E. (2004). Interpolation and extrapolation in human behavior and neural networks. *Journal of Cognitive Neuroscience*, **16**, 382-389.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, **62**, 255-262.

Hammond, K. R., & Stewart, T. R. (Eds.) (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. New York: Macmillan.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, **57**, 116-131.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximations. *Neural Networks*, **2**, 359-366.

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

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, **111**, 1072-1099.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 616-637.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 811-836.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionistic model of category learning. *Psychological Review*, **99**, 53-79.

Lambert, S. K., & Shanks, D. R. (Eds.) (1997). *Knowledge, concepts, and categories*. Hove, U.K.: Psychology Press.

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, **131**, 163-193.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press, Bradford Books.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316.

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

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**, 978-982.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**(Whole No. 517).

Uhl, C. N. (1963). Learning of interval concepts: I. Effects of differences in stimulus weights. *Journal of Experimental Psychology*, **66**, 264-273.

Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception & Psychophysics*, **18**, 416-422.

NOTES

1. For all models, no penalty was extracted for the constant and the first term (which reflects the linear component for the polynomial model and the fundamental frequency for the Fourier model). Additional terms were penalized proportionally to the order of the hypothesis for the polynomial model and Fourier model so that $\lambda_j = \lambda \cdot (j/n)$ for $j > 2$. For the logistic model, $\lambda_j = \lambda$, for $j > 2$.

2. For all parsimony rule models, the intercept and first hidden rule (linear component for the polynomial model; fundamental frequency for the Fourier model) were always included in the model, independent of their magnitudes, but no other term was entered until its validity v_j was driven by error reduction to exceed a cutoff δ in magnitude.

3. The reader should focus on the order of difficulty implied by the model, because the absolute magnitudes are not directly comparable. When the model predictions in Table 1B are compared with the empirical data in Table 1A, it is important to note that the human data represent responses of a single subject on a single trial averaged across trials and subjects, whereas the model's predictions represent the expectation, or mean response, averaged over the probability distribution of responses. Thus, the variance of the model predictions is smaller than the variance of the human data for each condition.

4. The R^2 is negative here because the mean response fits better than the model for the quadratic data, and thus the total sum of squared deviations around the mean is less than the sum of squared deviations from model predictions ($TSS < SSE$).

5. These all appeared while the present article was under review.

6. The formal model of the knowledge-partitioning approach appeared (Kalish et al., 2004) while the present article was under review.

(*Continued on next page*)

APPENDIX
Weights Connecting Inputs to Hidden Units for the Rule-Based Models

Polynomial Rule

The weights, w_{ji}, for this model are defined by a ($n \times 101$) matrix. The first row is simply constant across the columns and is normalized to have unit length. Each subsequent row was constructed from an orthonormal polynomial: The jth row contains the 101 scores for the jth-order orthogonal polynomial, corresponding to x^j for $x = 0, .01, .02, \ldots, 1.0$ and $j = 0, 1, 2, \ldots, n$. For example, the first row contains the flat or zero-order orthogonal polynomial (corresponding to x^0), the second row contains the linear or first-order orthogonal polynomial (corresponding to x^1), the third row contains the quadratic or second-order orthogonal polynomial (corresponding to x^2), the fourth row contains the cubic or third-order orthogonal polynomial (corresponding to x^3), and so on. In this case, Q is simply an identity function, $Q(x) = x$, in Equation 2.

Fourier Rule

The weights, w_{ji}, for this model are defined by a ($n \times 101$) matrix. The first row is simply constant across the columns and is normalized to have unit length. Each subsequent pair of rows was constructed from a pair of orthonormal harmonics: $\sin[2 \cdot \pi \cdot (j/101)x]$ defines the first of each pair, and $\cos[2 \cdot \pi \cdot (j/101)x]$ defines the second of the pair for $x = 0, 1, 2, \ldots, 100$ and $j = 1, 2, \ldots, (n - 1)/2$. For example, the first pair of rows contains the fundamental frequency (corresponding to the frequency for $j = 1$), the second pair of rows contains the first harmonic (corresponding to the frequency for $j = 2$), the third pair of rows contains the second harmonic (corresponding to the frequency for $j = 3$), and so on. In this case, Q is simply an identity function, $Q(x) = x$, in Equation 2.

Logistic Rule

For this model, $Q(x) = (1 + e^{-x})^{-1}$ in Equation 2. In this case, the weights, w_{ji}, form a ($n \times 101$) matrix in which each element is defined as an affine transformation of the input node index: $w_{ji} = w_j \cdot (i/100) + \theta_j$. Each row contains 101 scores that form a linearly increasing function of the input node index i with a different slope w_j and intercept θ_j selected for each row. The first row is a special case in which the slope is fixed equal to zero and the intercept is fixed equal to a large value, so that the output of the first logistic unit H_1 always equals 1.0. For the remaining rows, the slopes and intercepts (w_j, θ_j) for each logistic node H_j are learned during training using back-propagation:

$$w_{ji}(t + 1) = w_{ji}(t) + \beta\, a_i(x)\, Q'(h_j)\, v_j(t)\, [z(t) - y(x)] + \gamma\, [w_{ji}(t) - w_{ji}(t - 1)],$$

where β is a learning rate parameter, $Q'(x)$ is the derivative of the nonlinear transformation $Q(x)$, and the last term represents the momentum term moderated by a parameter γ.

Comparing Rules Using a Common Architecture

Note that the only change in the model required to represent the polynomial and Fourier rule models was a change in the weights, w_{ji}, used to map inputs to the hidden units. Nothing else in the connectionist architecture is changed with this comparison. Furthermore, the weight matrices for both cases are orthogonal and normalized to unit length, which further equates the two representations. Thus, this representation provides a comparison that holds all architectural features constant except for the representation of the rules by the weights.

The logistic model was formed not only by a change in weights but also by the use of a logistic output function $Q(x)$. All other aspects of the cognitive architecture are the same as for the polynomial rule and Fourier rule models.

(Manuscript received April 17, 2003;
revision accepted for publication April 22, 2004.)