Journal of Experimental Psychology:
Learning, Memory, and Cognition
1997, Vol. 23, No. 4, 968–986

0278-7393/97/$3.00

Extrapolation: The Sine Qua Non for Abstraction in Function Learning

Edward L. DeLosh
Colorado State University

Jerome R. Busemeyer
Purdue University

Mark A. McDaniel
University of New Mexico

Abstraction was investigated by examining extrapolation behavior in a function-learning task. During training, participants associated stimulus and response magnitudes (in the form of horizontal bar lengths) that covaried according to a linear, exponential, or quadratic function. After training, novel stimulus magnitudes were presented as tests of extrapolation and interpolation. Participants extrapolated well beyond the range of learned responses, and their responses captured the general shape of the assigned functions, with some systematic deviations. Notable individual differences were observed, particularly in the quadratic condition. The number of unique stimulus–response pairs given during training (i.e., density) was also manipulated but did not affect training or transfer performance. Two rule-learning models, an associative-learning model, and a new hybrid model with associative learning and rule-based responding (extrapolation–association model [EXAM]) were evaluated with respect to the transfer data. EXAM best approximated the overall pattern of extrapolation performance.

Research on conceptual behavior has historically focused on category learning and the application of categorical knowledge. So dominant is this focus that the terms *concept* and *category* are often used interchangeably (e.g., see Bourne, 1966; Smith & Medin, 1981). It is useful to distinguish these two terms, however, because there are many types of concepts that can not be adequately characterized as categories (Busemeyer, McDaniel, & Byun, 1997; Estes, 1995). In general, concepts also pertain to causal variables (e.g., intelligence) and relationships between these variables (e.g., income is correlated with intelligence). This article investigates functions, which conceptualize the relationship between causal variables.

By definition, a function maps a set of input values (called the *domain* of the function) into a set of output values (called the *range* of the function), such that each input value is assigned only one output value. In function-learning situations, the range is composed of a continuous set of response magnitudes (e.g., predict a student's GPA on the basis of IQ scores). In category learning, on the other hand, outputs consist of discrete and nominal response categories (e.g., classify a student as normal vs. gifted on the basis of IQ scores). We call concept-learning tasks that involve a continuum of response magnitudes *function-learning tasks* and those with discrete and nominal response categories *category-learning tasks.*

A central issue in theoretical treatments of concept learning is whether conceptual behavior is based on associations between previously encountered instances and assigned responses (e.g., Brooks, 1978; Estes, 1986; Gluck & Bower, 1988a, 1988b; Hintzman, 1986; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984, 1986) or also involves the abstraction of a rule that represents the relation between instances and responses (Anderson, 1990; Ashby & Gott, 1988; Smith & Medin, 1981; Trabasso & Bower, 1968). Although this is still a hotly debated issue, the balance of evidence seems to support an associative-learning or exemplar-based approach to concepts (for a review, see Nosofsky, 1992; Nosofsky & Kruschke, 1992). The evidence stems primarily from experiments on category learning, however, and it remains to be seen whether an associative-learning or a rule-based approach provides a better explanation of conceptual behavior with functions (for a comparison of these approaches, see Busemeyer, Byun, DeLosh, & McDaniel, in press; Koh & Meyer, 1991).

To discriminate between associative-learning and rule-based accounts as they pertain to functions, we examined extrapolation behavior in the present study. Consider the common uses of different types of concepts. Categories are typically used to assign new stimuli from a given conceptual domain to previously learned response categories (e.g., if *1, 2, 3, 5,* and *7* are prime numbers, is *9* also a prime number?). Abstract rules, on the other hand, may be used to construct new responses in reaction to novel stimuli from the conceptual domain (e.g., if *1, 2, 3, 5,* and *7* are prime numbers, what

Edward L. DeLosh, Department of Psychology, Colorado State University; Jerome R. Busemeyer, Department of Psychological Sciences, Purdue University; Mark A. McDaniel, Department of Psychology, University of New Mexico.

This research was supported by National Institute of Mental Health Grant MH47126. Some of the data reported here were presented at the May 1992 meeting of Hoosier Mental Life in West Lafayette, IN, and the November 1992 meeting of the Psychonomic Society in St. Louis, MO. We thank Sangsup Choi for his preliminary work on the simulation modeling for this study.

Correspondence concerning this article should be addressed to Edward L. DeLosh, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523. Electronic mail may be sent via Internet to delosh@lamar.colostate.edu.

is the next prime number after 9?). The latter use of concepts is called *extrapolation.* As demonstrated in the following example, function-learning tasks are ideally suited to studying extrapolation behavior.[1]

One possible function, a quadratic function, is shown in Figure 1. Suppose a participant is trained with a set of stimulus values that lie inside the two vertical lines shown in the figure and learns the response values that lie inside the two horizontal lines. An *extrapolation test* is defined as the presentation of a novel stimulus value that lies outside the training domain, that is, outside of the two vertical lines in Figure 1 (whereas an *interpolation test* involves the presentation of a novel stimulus value that lies within the training domain). Responding on an extrapolation test trial with a response magnitude that is outside the training range (i.e., outside the two horizontal lines in Figure 1) is termed an *extrapolation response.*

To illustrate the theoretical implications of extrapolation behavior, Carroll (1963) outlined extreme versions of associative-learning versus rule-based models as they apply to the above example. An extreme form of an associative-learning model assumes that participants store each stimulus–response pair presented during training and, when a new stimulus is given, produce the response associated with the most similar training stimulus. This model does not allow novel responses to be generated; therefore, one would not expect participants to extrapolate at all by this account. Instead, the response generated on an extrapolation test should be equivalent to one of the responses learned during training (i.e., a trained response at the boundary of the trained response range). An extreme form of rule-based models, on the other hand, states that participants abstract the training rule itself and, when a new stimulus is given, generate a new response consistent with that rule. For the function shown in Figure 1, an extrapolation response would be of a smaller magnitude than those learned during training.

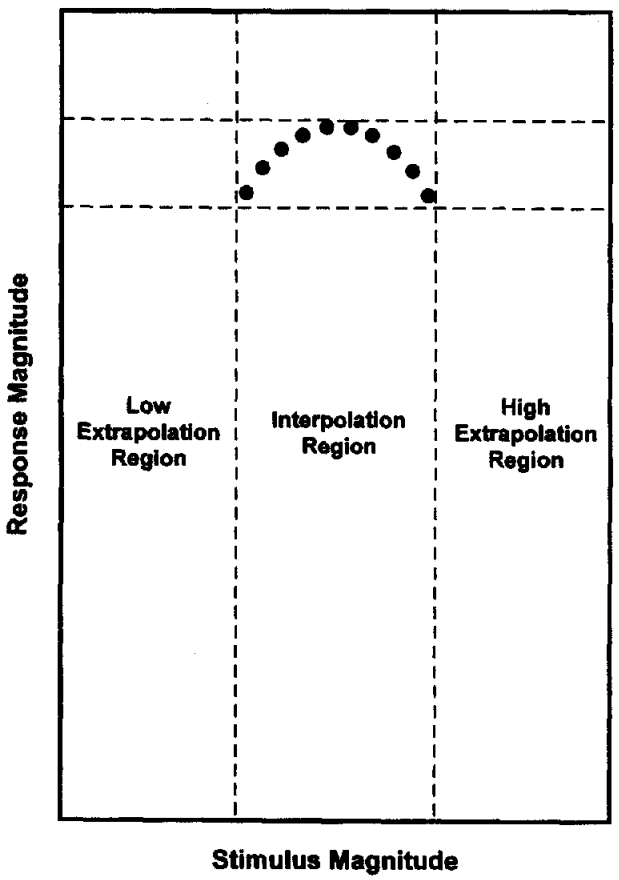

Figure 1. A quadratic functional relation between stimulus and response magnitudes, showing a sample stimulus–response set and the resultant interpolation and extrapolation regions.

Numerous studies have investigated the relative learning rate of different types of functions (e.g., Brehmer, 1974; Deane, Hammond, & Summers, 1972; Summers, Summers, & Karkau, 1969; for a review, see Busemeyer, Byun, et al., in press), but only three function-learning studies have investigated transfer to extrapolation tests. One is an unpublished technical report (Carroll, 1963), and another is a briefly mentioned experiment that was peripheral to the thrust of the chapter in which it was reported (Surber, 1987). The third study included three experiments in which participants had to predict future pollution levels when given pollution levels for the previous 5 years, but the scope of the study was limited to a single type of function (exponential) and a small range of extrapolation values (Waganaar & Sagaria, 1975). All three of these studies indicated that participants extrapolate in the direction of the training function, thereby suggesting that an extreme version of an associative model may be ruled out. Given the limited database available, this conclusion must be viewed as preliminary, however, and additional empirical research is warranted.

The need for further empirical work notwithstanding, the experiments cited above do not resolve the issue of whether functions are learned by associations or by rules because there are more sophisticated versions of associative-learning models that assume stimulus generalization (e.g., Kruschke, 1992; Nosofsky, 1984, 1986). When applying these models to function-learning tasks in which responses lie on a continuum, it is appropriate to include response generalization as well, and in this case some extrapolation is possible. Although the extent of extrapolation that can be generated by such models is still quite limited (as we show in a later section), the models remain viable because the extent to which humans extrapolate has not yet been established. Consequently, a more systematic investigation of extrapola-

[1] A common idea is that abstraction is necessary for extrapolation, but our idea, as implied in the title, is that extrapolation is the essential condition for abstraction. It is possible, for instance, that the need for accurate extrapolation may have led to the evolution of an abstraction process. The idea that extrapolation is a necessity for abstraction is especially appropriate in the context of this article, because in the extrapolation–association model (EXAM) that we propose, abstraction does not take place until extrapolation is called for. These are the deeper meanings of our title.

tion behavior vis-à-vis current models is needed. Toward this objective, we evaluated four different models of function learning with respect to extrapolation behavior.

We examined two rule-learning models: the polynomial hypothesis-testing model, first outlined by Carroll (1963) and later elaborated by Brehmer (1974), and the log-polynomial adaptive-regression model, proposed more recently by Koh and Meyer (1991). These rule-learning models were specifically developed for function-learning tasks to explain previous empirical findings concerning the learning rates of different types of functions. The third model we tested was the associative-learning model (ALM), recently proposed by Busemeyer, Byun, et al. (in press). It is an extension of an exemplar-based connectionist model called attention learning covering map (ALCOVE; Kruschke, 1992). ALCOVE was specifically developed for category-learning tasks to account for the learning rates of various category structures and is currently the most powerful formulation of this class (see Nosofsky & Kruschke, 1992). Note that ALCOVE was not developed with function-learning tasks in mind. However, Busemeyer, Byun, et al. showed that their extension of ALCOVE can account for the learning rates of different types of functions. Moreover, the stimulus and response generalization incorporated in this model allows for some extrapolation. ALM is therefore a viable alternative to rule-learning models, and it is worthwhile to determine whether this type of model—although originally developed for category learning—can be extended to account for extrapolation behavior in function learning.

The fourth model we tested was a new hybrid model called the extrapolation–association model (EXAM). This model is based on the same associative-learning assumptions as ALM, but it also incorporates a rule-based response mechanism capable of linear interpolation and extrapolation. EXAM is essentially an extension of ALM for producing systematic extrapolations beyond the range of learned responses, and it is motivated by Waganaar and Sagaria's (1975) observation that extrapolation is approximately linear, even for nonlinear (exponential) functions.

The present article is organized as follows. First, we present the empirical results of two experiments that investigated extrapolation in a function-learning task. We examined extrapolation for three function forms (linear, exponential, and quadratic) and three density conditions (low, medium, and high density, where density is the number of unique stimuli given during training). Then, a more complete description of the models is presented, followed by a comparison of the model predictions to the observed results. We conclude by summarizing the new findings on extrapolation behavior and evaluating associative versus rule-based explanations of function learning.

Experiment 1

In Experiment 1, we investigated three different function forms: a linear, an exponential, and a quadratic function. The linear function was included to evaluate extrapolation in the simplest possible case. If participants learned a linear rule, then their transfer responses should have continued to be linear in the extrapolation regions. The nonlinear functions were included to ensure that extrapolation, if found, was not limited to a simple linear relation.

Participants were first trained on stimuli from the middle of the range of possible stimulus magnitudes (i.e., inside the two vertical lines in Figure 1). After training, 45 transfer trials were presented as tests of extrapolation and interpolation. Three different types of transfer stimuli were used. Low-extrapolation items consisted of 15 new stimulus values sampled from the low end of the stimulus scale (to the left of the left vertical line in Figure 1); high-extrapolation items consisted of 15 new stimulus values sampled from the high end of the stimulus scale (to the right of the right vertical line in Figure 1); and interpolation items consisted of 15 new stimulus values sampled from the middle of the scale (inside the two vertical lines and of different values than the training stimuli). Although we focused on extrapolation performance, we included interpolation trials in order to compare directly performance for novel stimuli from within the training domain to performance for those that lie outside the training domain. Of foremost concern was the extent to which participants extrapolated beyond learned responses and the nature of their extrapolations (e.g., would they overestimate or underestimate the training function?).

In order to evaluate the generality of the results, we also manipulated the density of training stimulus magnitudes. In a low-density condition, participants were trained with 25 replications of 8 equally spaced stimulus values. In a medium-density condition, participants received 10 replications of 20 stimulus values. In a high-density condition, participants received 4 replications of 50 stimulus values. In each case, the set of training stimuli spanned the entire training domain. According to an associative-learning view, the function-learning task used was equivalent to learning a list of stimulus–response associations. Therefore, the high-density condition corresponded to a long stimulus list, and the low-density condition corresponded to a short stimulus list. The typical finding in associative learning is that longer lists are learned more slowly than short lists (e.g., Gillund & Shiffrin, 1984; Murdock, 1962; Roberts, 1972; Waugh, 1972). Thus, an associative-learning framework led us to expect slower learning in the high-density condition than the low-density condition.

The density manipulation might also have affected transfer performance in the following manner. Participants may have relied on stimulus–response associations (i.e., memory for exemplars) when given a small set of stimulus values repeated many times (cf. Homa, Sterling, & Trepel, 1981; Kellogg & Bourne, 1989); therefore, they may have exhibited little extrapolation. When given a large set of stimulus–response pairs, participants may instead have used a rule-based strategy to summarize the mapping between stimuli and responses, thereby promoting extrapolation. If density did influence strategy choice in this fashion, extrapolation performance would be better for the high-density condition than the low-density condition.

Method

Participants and apparatus. One hundred and eight Purdue University undergraduates participated in partial fulfillment of a requirement for an introductory psychology course. The experiment was conducted in a small laboratory room equipped with a single desk and microcomputer. Participants sat at the desk and viewed the experiment on a 14-in. color monitor from a distance of about 60 cm; they responded by using a standard keyboard placed on the desk in front of the monitor. A computer program controlled the presentation of the instructions and stimuli and collected participant responses.

Design. The experiment used a 3 × 3 (Function × Density) between-subjects factorial design with 12 participants randomly assigned to each of the nine experimental conditions. Density, defined as the number of unique stimulus magnitudes (inputs) presented during training, was either 8 (low), 20 (medium), or 50 (high). The correct response magnitudes (outputs) corresponding to the stimulus magnitudes were computed using either a linear, exponential, or quadratic function.

The training phase of the experiment consisted of 200 correct-response feedback trials. The scale of possible stimulus magnitudes ranged from 0 to 100 (as indicated by an unfilled horizontal bar labeled 0 to 100 in 10-point increments). However, to allow for extrapolation trials during the transfer phase of the experiment, the magnitudes presented during training were limited to values between 30 and 70. Within the specified input range of 30 to 70, participants were presented either 8, 20, or 50 unique stimuli according to the assigned density level. A single stimulus set was constructed for each density condition, such that the stimuli were spaced out as evenly as possible within the training domain, given the constraint of rounding off to half-point values and the necessity of excluding points to be used on interpolation trials (see Appendix A for the exact stimulus magnitudes). During the training phase, the assigned stimulus set was presented as a block of trials, such that each input magnitude was presented once and only once within a trial block. These blocks were then repeated until a total of 200 trials were presented. Thus, there were 25, 10, and 4 block repetitions (and therefore item repetitions) for the low-, medium-, and high-density conditions, respectively. The order of stimulus presentation within a trial block was randomized separately for each block and each participant.

The response magnitudes used for feedback during the training phase were computed using the following equations: $y = 2.2x + 30$ for the linear function; $y = 200(1 - e^{-x/25})$ for the exponential function; and $y = 210 - (x - 50)^2/12$ for the quadratic function. The range of response magnitudes allowed on training and blank trials was 0 to 250 (as indicated by an unfilled horizontal bar labeled 0 to 250 in 10-point increments). The outputs produced by the assigned function were rounded to the nearest integer prior to screen representation.

The transfer phase of the experiment consisted of 45 trials without feedback, using stimulus magnitudes that had not been presented during training. Fifteen equally spaced values were selected from the training domain for interpolation trials; 15 values from below the training domain were selected for low-extrapolation trials; and 15 values from above the training domain were selected for high-extrapolation trials (see Appendix B). Identical transfer stimuli were used in all conditions of the experiment. During the transfer phase, the 15 stimulus magnitudes from each region were grouped together as a trial block. The order of presentation of the three resulting trial blocks was counterbalanced across conditions. In addition, the order of stimulus presentation within a block was randomized separately for each block and each participant.

Procedure. At the beginning of the experiment, participants read two pages of instructions on the computer monitor. In these instructions, participants were told that they were to learn by trial and error the relation between amounts of an unknown substance and the levels of arousal they cause in humans, using feedback as a guide. The instructions also described the format of the presentation screen and the appropriate keys to use for making a prediction. Participants were not informed that they would be given new trials after training. Once these instructions were understood, a sample trial was provided in order to familiarize the participant with the presentation screen and response procedure.

After the sample trial, participants proceeded with the training phase of the experiment. During training, three unfilled horizontal bars were presented simultaneously on the monitor. The top bar was titled Substance X and had tick marks and value labels every 10 units from 0 to 100; the remaining two bars were titled Predicted Level of Arousal and Actual Level of Arousal, respectively, with tick marks and value labels every 10 units from 0 to 250. The relative lengths of these unfilled bars on the screen were proportional to the number of units they represented. On a given trial, the uppermost bar was filled in from the zero point (at the left end of the bar) to the input value representing the amount of substance administered. Participants then used the arrow keys to fill in the second bar from the zero point to the desired prediction value and pressed the space bar when finished. Once the space bar was pressed, the correct level of arousal (i.e., the output value of the assigned function) was represented on the third horizontal bar. Participants were also shown the absolute deviation of their prediction from the correct response and an accuracy score that ranged from 0 to 100, which was computed as 100 minus the square of the deviation. The next trial was initiated when the participant pressed the enter key. The time allotted for making a prediction and for studying feedback was participant-determined, although participants were told that they should not spend more than 20 s on any one trial in order to complete the experiment on time.

Once participants completed the training phase of the experiment, they were given a 1-min break, which was followed by transfer instructions and the transfer tests. Transfer trials proceeded exactly like training trials except the output bar was not displayed and feedback was no longer provided.

Results and Discussion

Training performance. The measure used to examine training performance was the absolute deviation of participants' predictions from the correct function value for each training trial. The absolute deviations were averaged over blocks of 20 trials, yielding 10 successive average deviation scores for each participant.[2] An alpha level of .05 was used for this and all other analyses reported in this study.

[2] Participants occasionally responded with the default value of zero or produced responses highly inconsistent with their trend of predictions by accidentally hitting the space bar. Four participants also made a few erratic predictions toward the end of training. Therefore, a five standard deviation rule was invoked in an attempt to filter out extreme outliers. Any prediction whose error (deviation from the assigned function) was more than five standard deviations greater than the average error of the preceding block of 20 trials was eliminated. Using this method, we removed less than 0.5% of the data points, with no apparent bias toward any function or density condition.

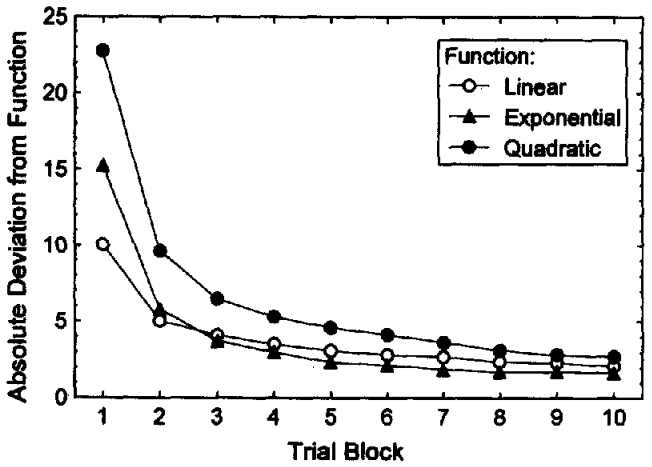

Figure 2. The mean deviation of participants' responses from the assigned linear, exponential, and quadratic functions across blocks of training trials in Experiment 1.

Figure 2 shows the average absolute deviation for each trial block and training function. At the beginning of training, performance was best for the linear function, second best for the exponential function, and worst for the quadratic function. As training progressed, this difference diminished: The learning curves converged to an average absolute deviation of 2.4 units on a response scale of 250 units. A 3 × 3 × 10 (Function × Density × Trial Block) mixed analysis of variance (ANOVA) supported this impression, yielding a Function × Trial Block interaction, $F(18, 891) = 22.25$, $MSE = 5.67$, $p < .001$. These relative learning rates are consistent with past research on the learnability of functions, which has shown that monotonic functions are learned faster than nonmonotonic functions (e.g., Brehmer, 1974; Carroll, 1963; Deane, Hammond, & Summers, 1972) and that linear functions are learned slightly faster than nonlinear monotonic functions (see Busemeyer, Byun, et al., in press, for a review). Thus, the current training procedures reproduced standard findings.

It was somewhat surprising that there were no significant effects of density on training performance ($F < 2.40, p > .10$, for the main effect and interactions).[3] A post hoc explanation is that participants in the low-density condition benefited from more item repetitions, and participants in the high-density condition benefited from an increase in interitem similarity (whereas participants in the medium-density condition benefited from both to a lesser extent). Thus, the lack of differences in training performance as a function of stimulus density may have resulted from concurrent effects of repetition and similarity.

Transfer performance. The three panels of Figure 3 show the average of participants' predictions plotted as a function of the stimulus input values given during transfer. The left, center, and right panels display the results for the linear, exponential, and quadratic functions, respectively. Each panel is divided into three regions showing the participants' predictions in the low-extrapolation, interpolation, and high-extrapolation regions.

First, consider the interpolation region. It is apparent that participants, on average, approximated the assigned functions very well in this region. The average absolute deviation score for interpolation was 3.2, which compares favorably with the 2.4 deviation score achieved at the completion of training. Participants performed nearly as well on the interpolation transfer stimuli as they did on the trained stimuli, despite the fact that the transfer stimuli had not been shown previously.

Next, consider the extrapolation regions. Participants in all three function conditions extrapolated much beyond the range of response magnitudes learned during training. In addition, the pattern of responses produced by participants captured the general shape of the functions in the extrapolation regions even though participants were trained on a considerably limited range of stimulus magnitudes. This general finding poses problems for possible associative-learning accounts of function learning, because extant associative models do not provide a mechanism for extrapolating far outside the range of learned responses (as we demonstrate in a later section).

Another aspect of the findings that is noteworthy is that participants' predictions deviated more from the assigned function in the extrapolation regions than in the interpolation region. Furthermore, and perhaps of greater importance, participants' predictions underestimated the assigned function for the linear condition but overestimated the assigned function for the exponential and quadratic conditions. DeLosh (1995) has replicated these results for the linear and quadratic functions (also see Byun, 1995). This pattern of over- and underestimation poses problems for rule-based models of function learning. If participants abstracted a simple rule for the linear condition, for example, then they should perform equally well on extrapolation and interpolation tests. However, participants systematically underestimated the linear function in both extrapolation regions even though they were very accurate on interpolation trials. Thus, the ensemble of transfer results challenges both associative-learning and rule-based models as they apply to function

[3] An ANOVA using data that had not been trimmed with the five standard deviation rule yielded the same results given above. Several additional analyses of the training data were also conducted: one in which average deviation scores that increased at the end of training (which occurred in 4 participants) were replaced with the average deviation score of the preceding trial block; another in which data from the first 100 trials were examined; and yet another in which data from the 6 most accurate participants from each condition were tested. The only new findings occurred in the latter analysis, which yielded a density main effect ($p < .01$), qualified by a Function × Density interaction ($p < .05$). An analysis of simple effects revealed that the effect of density on training performance was localized to the linear function, where the deviation scores were 1.9, 3.1, and 5.2 for the density conditions of 8, 20, and 50, respectively ($p < .001$). Thus, only among the very best learners was a substantial effect of density on training evident, and among these participants the effect occurred only in the case of the linear function.

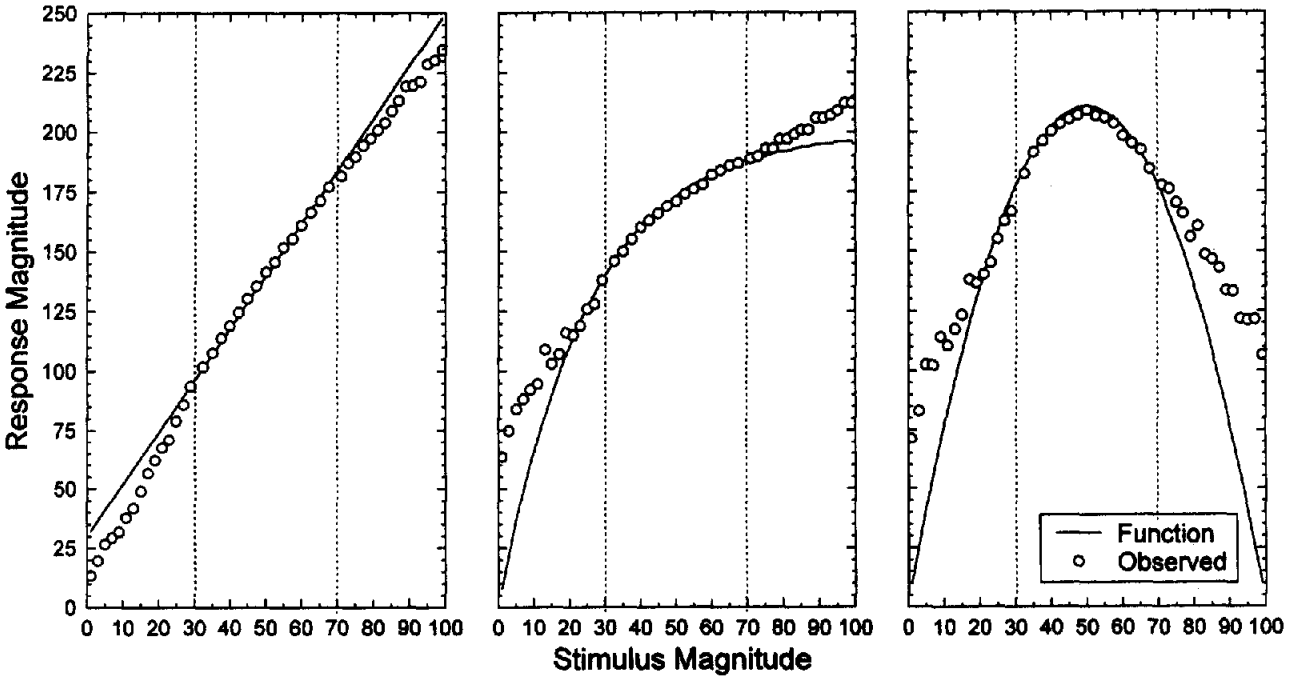

Figure 3. The mean of participants' predictions across transfer trials for the linear, exponential, and quadratic functions in Experiment 1.

learning. This issue is considered more rigorously in a later section where we formally test the models mentioned earlier.

We conducted a statistical examination of the transfer data by performing a 3 × 3 × 3 (Function × Density × Transfer Region) mixed ANOVA, using the signed deviations of participants' predictions from the assigned function.[4] This analysis yielded main effects of both function and region, $F(2, 99) = 34.43$, $MSE = 562.72$, $p < .001$, and $F(2, 198) = 25.16$, $MSE = 298.03$, $p < .001$, respectively, qualified by a Function × Region interaction, $F(4, 198) = 27.41$, $MSE = 298.03$, $p < .001$. Analyses of simple effects revealed that performance did not differ across function conditions in the interpolation region, $F(2, 269) = 0.07$, $MSE = 386.26$, $p > .10$, but did significantly differ across functions in the low- and high-extrapolation regions, Fs(2, 269) = 31.25 and 61.15, respectively, MSEs = 386.26, ps < .001, supporting the general impressions discussed above. The density manipulation did not reliably affect transfer performance (Fs < 2.30, ps > .05, for the main effect and interactions).

Individual differences. To this point, we have focused our attention on group performance. In order to get a sense of the degree to which group data were representative of individual learners rather than an artifact of averaging over learners, we examined the learning and transfer data for individual participants. For the linear and exponential functions, all participants extrapolated in the direction of the group averages, indicating that the group data were reflective of individual processes. For the quadratic group, most participants performed in accordance with the group average, but there were seven notable exceptions. Therefore, we describe individual differences for the quadratic condition in greater detail.

Of the 36 participants, 29 extrapolated at both ends of the function, as seen in the group data; Figure 4 reveals that at least some of them approximated the quadratic function very well in their extrapolations. Five of the 36 participants extrapolated in the low-extrapolation region but not the high-extrapolation region; Figure 5 shows two examples of this type of pattern. Two participants did not extrapolate in either region; Figure 6 shows the response patterns for these participants.[5] The latter 2 participants behaved in a manner

[4] Signed deviations were analyzed instead of absolute deviations to test statistically for the over- and underestimation patterns observed in Figure 3. As with the training data, outliers in the transfer data were excluded using a five standard deviation rule. Any prediction with an error more than five standard deviations larger than the average error of the appropriate transfer region (interpolation, low extrapolation, or high extrapolation) was eliminated. In all, less than 0.3% of transfer responses were excluded.

[5] In order to separate participants into qualitative groups, the following value was computed for each participant and extrapolation region:

$$z = \frac{X^e - X^t}{s^i},$$

where X^e is the mean of the participant's predictions from the five most extreme extrapolation input values in the extrapolation region, X^t is the output value at the boundary of the training range, and s^i is the standard deviation of that participant's error in the interpolation region. A score of 3 was selected as the criterion as to whether a participant extrapolated in an extrapolation region. This criterion value was not critical because there were no borderline cases. All participants either did not exceed the learned response range in absolute value or did so with a z score of at least 10.54.

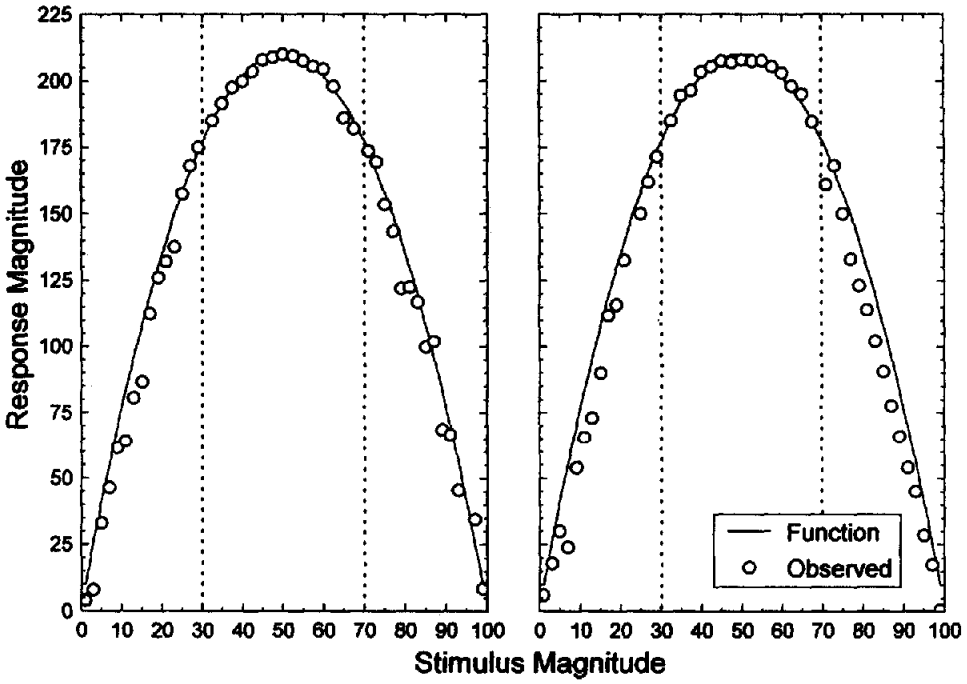

Figure 4. The transfer predictions of 2 individual participants who closely approximated the assigned quadratic function.

consistent with a simple associative model. Their responses on neighboring stimulus values were very similar, yielding flat segments in the response curves, which are especially striking in the left panel. Moreover, the output magnitudes of these flat segments corresponded to the responses learned during training, indicating that these participants produced old training responses to new transfer stimuli. Thus, it appears that there were at least three types of learners as typified by their extrapolation performance. Note that if we examined only training or interpolation performance, these differences among individual learners would not have been revealed.[6]

Experiment 2

Because very little empirical data exist on extrapolation in function-learning tasks, we conducted a second experiment to examine the reliability of our findings and to assess the generalizability of the results over variations in the procedural details of the task. Experiment 2 generally replicated the quadratic function condition of Experiment 1, but two potentially important aspects of the design and procedure were altered. In the initial instructions, participants were not told to learn the stimulus–response relationship, nor were they told that such a relationship was present. This change was incorporated to determine whether the results of Experiment 1 would generalize to a condition in which participants were not explicitly instructed to attend to the relationship between stimulus and response dimensions. In addition, the particular stimulus–response interpretation was changed to amount of growth hormone and plant height to ensure that the extrapolation results in Experiment 1 did not stem from prior knowledge about arousal functions.

Method

Twenty-four Purdue University undergraduates participated for pay; their pay ranged from \$4 to \$7, depending on the accuracy of their performance. Several aspects of the design and procedure were modified from Experiment 1 to assess the generalizability of the results over variations in procedural details. First, participants were instructed simply to learn each individual stimulus–response pair as it was presented. They were not told to determine the relationship between stimuli and responses, nor were they told that the stimuli and responses were in any way related. Second, the stimulus and response interpretation (i.e., cover story) involved the amount of a plant hormone and plant height, rather than drug dosage and arousal. Third, the assigned function was defined as $y = 230 - (x - 50)^2/12$, differing from that of Experiment 1 by a constant of 20 units, and training consisted of three replications of 20 unique stimulus values. Finally, nonfeedback test trials were presented systematically during the course of training. This last feature is not relevant for the current purposes and is not discussed further. Experiment 2 was identical to Experiment 1 in all other respects.

[6] Note that when the nonextrapolators were no longer included in the group data, participants' predictions still overestimated the assigned quadratic function in the extrapolation regions. Excluding the 5 participants who did not extrapolate in the high region, average predictions extended to 71.7 and 88.7 at the low and high extremes of the function, respectively. Excluding these 5 participants and the 2 who did not extrapolate in the low or high region, predictions extended to 62.6 and 81.0 at the low and high extremes of the function, respectively.

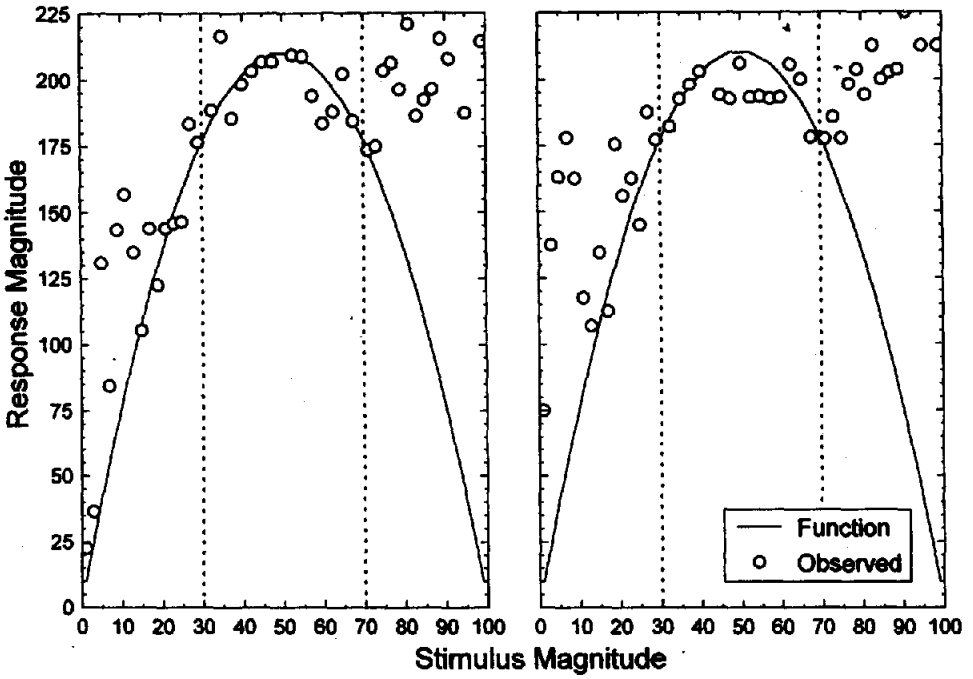

Figure 5. The transfer predictions of 2 individual participants in the quadratic function condition who did not extrapolate in the high-extrapolation region.

Results and Discussion

The average deviation from the assigned function for the last 10 trials of training ($M = 4.95$) was very similar to that obtained in Experiment 1 after the same number of trials ($M = 4.79$). The average of participants' predictions across the 45 extrapolation and interpolation trials given in Experiment 2 are shown in Figure 7. As before, interpolation predictions closely approximated the quadratic function. More important, extrapolation performance was nearly

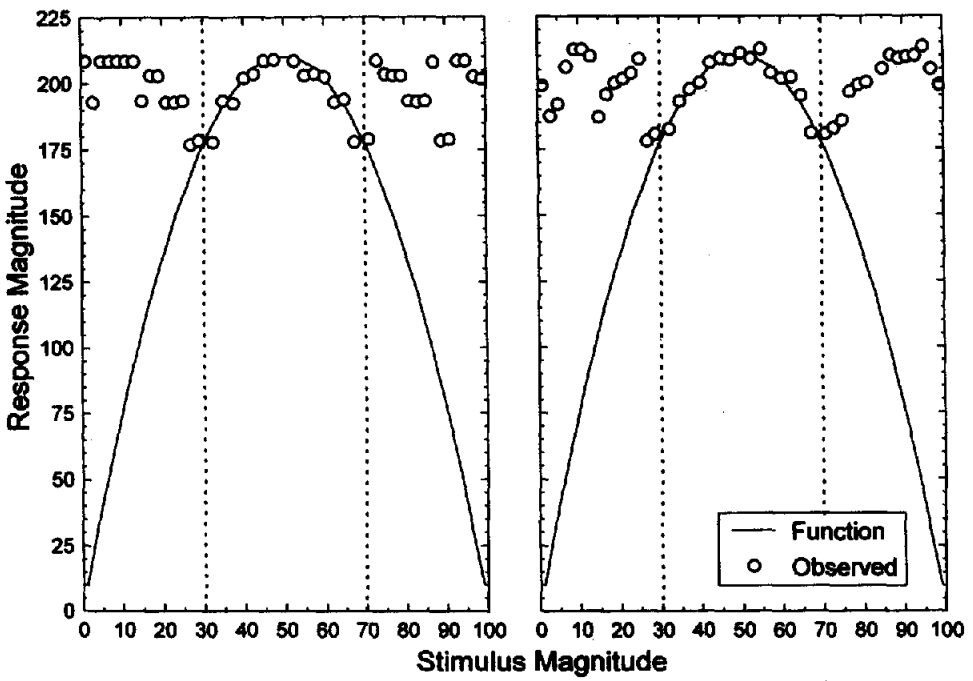

Figure 6. The transfer predictions of individual participants in the quadratic function condition who did not extrapolate in either extrapolation region.

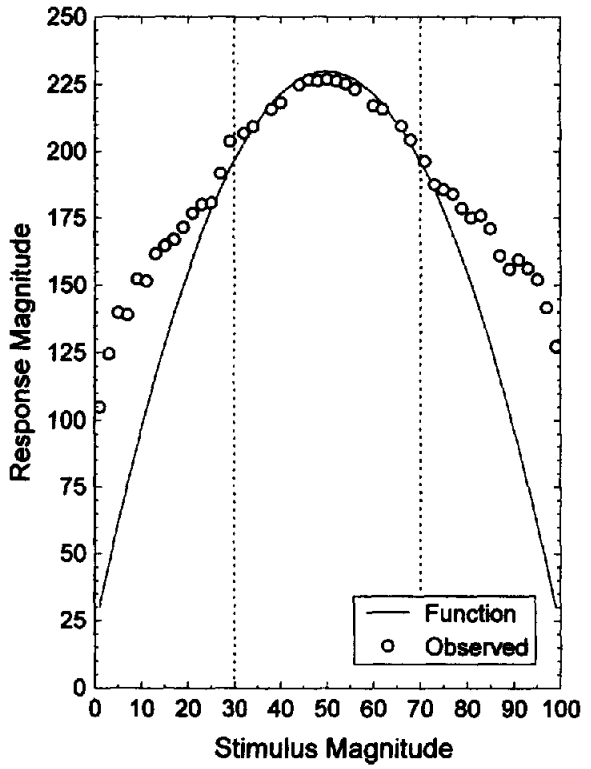

Figure 7. The mean of participants' predictions across transfer trials in Experiment 2.

identical to that observed in Experiment 1: Participants extrapolated well beyond learned responses, yet their extrapolations overestimated the assigned function in both extrapolation regions. Moreover, the extent of overestimation was comparable to that observed previously. Therefore, the pattern of transfer performance obtained in the first experiment did not require instructions that encouraged the induction of functional relations and was not particular to the stimulus–response interpretation used in Experiment 1.

Evaluation of Learning Models

In this section, the four different models of function learning mentioned previously are evaluated with respect to the data from Experiment 1. First, we describe the general method used to estimate parameters and test the competing models, and then we give a detailed description and evaluation of each model.

General Model-Testing Procedure

Each of the four models has two or three unknown parameters that must be estimated from the data. We estimated these model parameters separately for each function condition (linear, exponential, and quadratic) by searching for values that produced accurate fits to the last 50 trials of training. The last 50 trials of training were used because (a) the parameters were used to generate predictions on subsequent transfer tests and (b) these trials were less affected by the particular stimulus sequence used to train each participant, which was randomized to enhance the generality of the empirical results. The conclusions discussed in this section did not change, however, when parameter estimation was based on all 200 training trials or when the parameters were fit directly to the interpolation data.

For function condition k (k = linear, exponential, or quadratic), we chose model parameters that minimized the mean absolute error (*MAE*),

$$MAE_k = [\Sigma_j \Sigma_{t=151,200} |R_{jk}(t) - Y_{jk}(t)|]/(3) \cdot (50),$$

where $R_{jk}(t)$ represents the mean response (averaged across 12 participants) to the stimulus presented on trial t ($t = 151, \ldots, 200$) for density condition j ($j = 8, 20$, or 50) and function condition k, and Y_{jk} represents the model prediction for density condition j and function condition k. Each model was required to produce a fit within 1.2 units of *MAE*, which approximated the mean deviation of R_k from the programmed function at the end of training. The estimated parameters were then used to generate model predictions for the transfer trials. Note that with this method, no new model parameters were estimated from the transfer data, and this provides the critical test of the competing models.

Polynomial Hypothesis-Testing Model

If it is assumed that a single rule is learned to map stimulus magnitudes (X) to response magnitudes (Y), then this rule must be sufficiently general or flexible to accommodate a wide variety of mappings. The following polynomial regression model satisfies this property:

$$Y = b_0 + b_1 \cdot X + b_2 \cdot X^2 + b_3 \cdot X^3 + \cdots + b_p \cdot X^p. \quad (1)$$

For large p, a wide variety of mappings can be closely approximated by this polynomial.

According to the polynomial hypothesis-testing model, Equation 1 is used to generate responses, and the coefficients ($b_0, b_1, \ldots, b_p$) of this equation are learned during training. Prior hypotheses about the form of the mapping can be incorporated by assuming that the learning process begins with a particular set of coefficients. For instance, the learning process may begin with a linear hypothesis by initially setting $b_2 = b_3 = \ldots = b_p = 0$.

A trial-by-trial learning algorithm for learning the coefficients was also needed. The previous developers of the polynomial hypothesis-testing model (Brehmer, 1974; Carroll, 1963) did not propose any specific algorithm; therefore, we borrowed a standard learning algorithm for sequential estimation of regression coefficients that has been used in engineering (Ljung & Soderstrom, 1983). The $1 \times (p + 1)$ row vector, $\mathbf{b}' = [b_0, b_1, b_2, \ldots, b_p]$, represents the set of

coefficients that are to be learned; the $1 \times (p + 1)$ row vector, $\mathbf{X}' = [X^0, X^1, X^2, \ldots, X^p]$, represents the values of the polynomial terms in Equation 1; the symbol $Y(t)$ denotes the model prediction produced on trial t; and $Z(t)$ denotes the feedback signal on trial t. The learning algorithm is given in the following set of equations, where $\mathbf{I}$ is a $(p + 1)$ dimensional identity matrix, and $\mathbf{P}(1) = \mathbf{I}$:

$$\mathbf{b}(t) = \mathbf{b}(t - 1) + \alpha \cdot D(t)$$

$$D(t) = \mathbf{P}(t)/[1 + \mathbf{X}(t)'\mathbf{P}(t)\mathbf{X}(t)] \cdot [Z(t) - Y(t)] \cdot \mathbf{X}(t)$$

$$\mathbf{P}(t + 1)$$

$$= \mathbf{P}(t) \cdot [\mathbf{I} - \mathbf{X}(t)\mathbf{X}(t)'\mathbf{P}(t)]/[1 + \mathbf{X}(t)'\mathbf{P}(t)\mathbf{X}(t)]. \quad (2)$$

This learning rule is very similar to the delta learning algorithm (Rumelhart & McClelland, 1985). The only difference is the inclusion of the term, $\mathbf{P}(t)/[1 + \mathbf{X}(t)'\mathbf{P}(t)\mathbf{X}(t)]$. If this term is replaced with $\mathbf{I}$, the identity matrix, then Equation 2 is identical to the delta rule. The new term is essential for polynomial models because of differences in the scales of the polynomial terms and large correlations among these terms. It also increases the rate of learning, and theorems have been proven to show that the algorithm converges with training on the set of coefficients that minimizes the mean squared prediction error (Ljung & Soderstrom, 1983).

The polynomial hypothesis-testing model has two parameters, the power of the polynomial (p) and the learning rate (α). These parameters were estimated separately for each function condition k by searching for the lowest power and the smallest learning rate that would produce a MAE_k less than 1.2. Note that lower powers are preferred because high-order polynomials produce undesirable nonmonotonic oscillations in the extrapolation region. Table 1 shows the results of fitting this model to the learning data. The three rows labeled "polynomial" give the estimated parameters and *MAE* values of the polynomial model for the three function conditions. As expected, for the linear condition, a linear model ($p = 1$) was sufficient to fit the asymptotic learning data within a *MAE* of 1.2; for the quadratic condition, a quadratic model ($p = 2$) was sufficient. For the exponential condition, a cubic model ($p = 3$) was needed.

Panels A of Figures 8, 9, and 10 show the predictions of the polynomial hypothesis-testing model for each function condition across the 45 transfer trials by using the parameters estimated from the training data. Note that the model reproduced the assigned functions in the interpolation region, which matched the accurate responses of the participants in this region. The polynomial model also reproduced the assigned function in the extrapolation regions for the linear and quadratic conditions, failing to match the over- and underestimation pattern of responses produced by participants. In the case of the exponential function, the polynomial model overestimated the function in the high-extrapolation region but reproduced the function in the low-extrapolation region, unlike participants who overestimated the function in both extrapolation regions. We examined values of the polynomial parameter from $p = 1$ to $p = 10$, and no value of p yielded the observed pattern of over- and underestimation. In addition, the fit of the model was insensitive to changes in the learning rate for $\alpha > 1$. Thus, the polynomial model generally failed to explain the extrapolation results.

Table 1
Mean Absolute Error (MAE) and Estimated Parameter Values for Each Model and Function Condition

Model and function	*MAE*	Parameter values		
		α	*p*	γ
Polynomial				
Linear	0.95	5.0	1	
Exponential	0.75	1.1	3	
Quadratic	1.12	1.1	2	
Log polynomial				
Linear	0.95	1.0	5	
Exponential	0.86	3.0	2	
Quadratic	1.14	1.2	8	
ALM				
Linear	1.07	0.8		2
Exponential	0.76	0.9		3
Quadratic	1.07	0.9		3
EXAM				
Linear	1.07	0.8		2
Exponential	0.76	0.9		3
Quadratic	1.07	0.9		3

Note. α = learning rate; *p* = power of the polynomial; γ = scaling parameter; ALM = associative-learning model; EXAM = extrapolation–association model.

Log-Polynomial Adaptive-Regression Model

The polynomial model discussed above is just one possible implementation of a general and flexible rule. Koh and Meyer (1991) proposed the following alternative:

$$S = \ln(X), V = \ln(Y)$$

$$V = b_0 + b_1 \cdot S + b_2 \cdot S^2 + b_3 \cdot S^3 + \cdots + b_p \cdot S^p. \quad (3)$$

One justification for this model is that the physical stimulus (X or Y) has a nonlinear relation to its subjective image (S or V), and that it is most appropriate to base the learning rule on subjective images.

According to Koh and Meyer's (1991) model, responses are generated using Equation 3, and the coefficients ($b_0, b_1, \ldots, b_p$) of the polynomial are learned during training so as to minimize the following loss function:

$$L = \lambda \cdot L_1 + (1 - \lambda) \cdot L_2, \text{ where}$$

$$L_1 = \Sigma_{t=1,N}[V(t) - \ln Z(t)]^2,$$

$$L_2 = \int [\Sigma_{j=2,p} j \cdot (j - 1) \cdot b_j \cdot S^{j-2}]^2 \delta S, \quad (4)$$

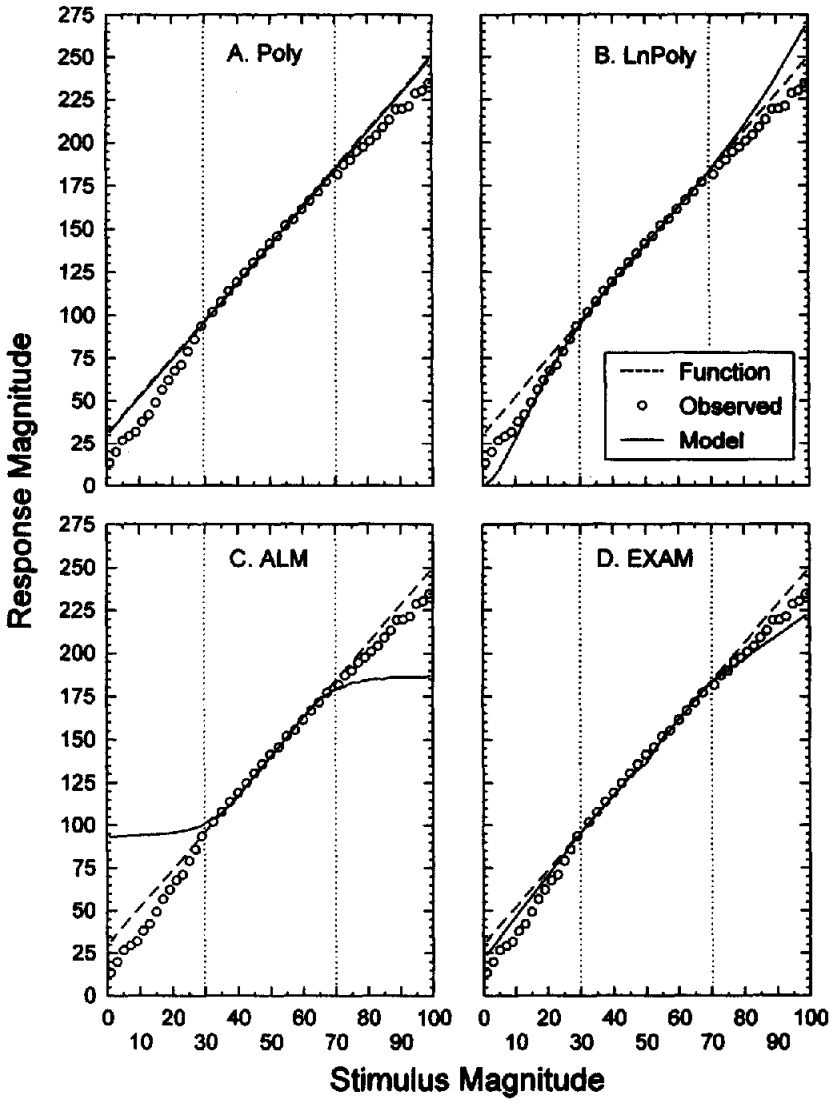

Figure 8. Model predictions and observed data for the linear function. Poly = polynomial hypothesis-testing model; LnPoly = log-polynomial adaptive-regression model; ALM = associative-learning model; EXAM = extrapolation–association model.

and ln $Z(t)$ is the log of the feedback signal on trial t. The first component (L_1) is a measure of accuracy (sum of squared prediction error), and the second component (L_2) is a measure of parsimony (a curvature index). The curvature index is only effective at the beginning of training (small N) and forces the model to begin with a simple function form. Later in training, the accuracy component dominates.

Koh and Meyer (1991) did not propose a specific trial-by-trial learning algorithm; therefore, we used the following:

$$\mathbf{b}(t) = \mathbf{b}(t-1) + \alpha \cdot [\lambda \cdot D(t) + (1-\lambda) \cdot \Delta L_2/t], \quad (5)$$

where $D(t)$ is the same as in Equation 2, except that X, Y, and Z are replaced with ln$X$, ln$Y$, and ln$Z$, respectively, and Δ is the gradient of L_2. The first component $D(t)$ changes the coefficients in the direction of minimizing the sum of squared prediction error. The second component is the negative gradient of the parsimony index, which changes the coefficients in the direction of a simpler model.

This model has three parameters: the highest power of the polynomial (p), the learning rate (α), and the weight given to accuracy versus parsimony (λ). These parameters were estimated separately for each function condition k by searching for the lowest power and the smallest learning rate that would produce a MAE_k less than 1.2. The three rows labeled "log polynomial" in Table 1 show the estimated parameters and MAE values of the log-polynomial adaptive-regression model for the three function conditions. Note that in every case the best fitting value of λ was approximately zero, so this parameter is not shown in the table. For the

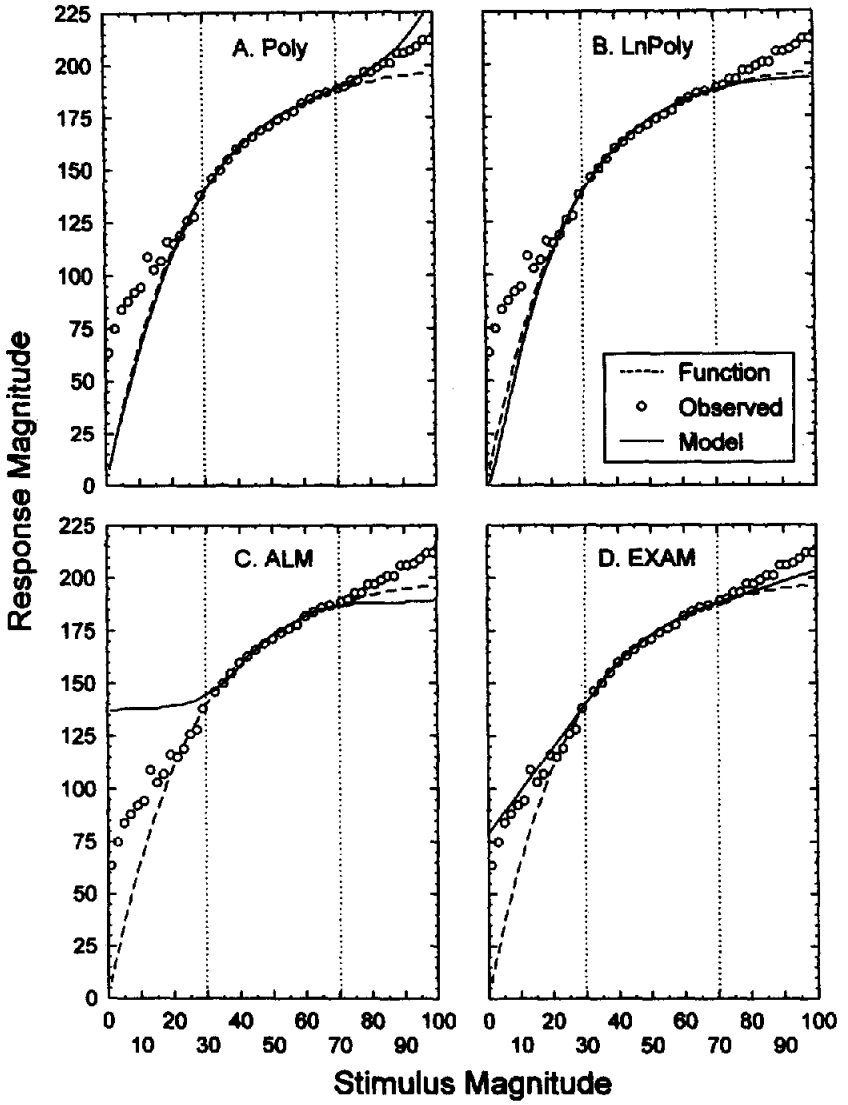

Figure 9. Model predictions and observed data for the exponential function. Poly = polynomial hypothesis-testing model; LnPoly = log-polynomial adaptive-regression model; ALM = associative-learning model; EXAM = extrapolation–association model.

linear condition, a fifth-order ($p = 5$) model was needed, because a simple line in (X, Z) coordinates is a nonlinear curve in $(\ln X, \ln Z)$ coordinates. For the quadratic condition, a quadratic model ($p = 2$) was sufficient to fit the asymptotic learning data, and for the exponential condition, an eighth-order ($p = 8$) model was needed.

Panels B of Figures 8, 9, and 10 show the predictions of the log-polynomial adaptive-regression model for each function condition across the 45 transfer trials, using the parameters estimated from the training data. The model reproduced the assigned function in the interpolation region, which matched the accurate responses of participants in this region. The log-polynomial model did not reproduce the assigned functions in the extrapolation regions, however, and of course neither did participants. But the over- and underestimation pattern predicted by the model did not correspond to the observed pattern. No value of the power parameter from $p = 1$ through $p = 10$ duplicated the over- and underestimation pattern of participants, and the fit of the model was insensitive to the changes in the learning rate for $\alpha > 1$. Thus, the log-polynomial adaptive-regression model also failed to explain the extrapolation results.

Associative-Learning Model (ALM)

According to ALM, the mapping between a set of stimuli and a set of responses is learned by associating M input nodes $\{X_1, X_2, \ldots, X_i, \ldots, X_M\}$ to L output nodes $\{Y_1, Y_2, \ldots, Y_i, \ldots, Y_L\}$. Each input node X_i corresponds to a position on the real number line that is proportional to one of the possible stimulus magnitudes. In the present experiments, the inputs ranged from 0 to 100 units in half-unit

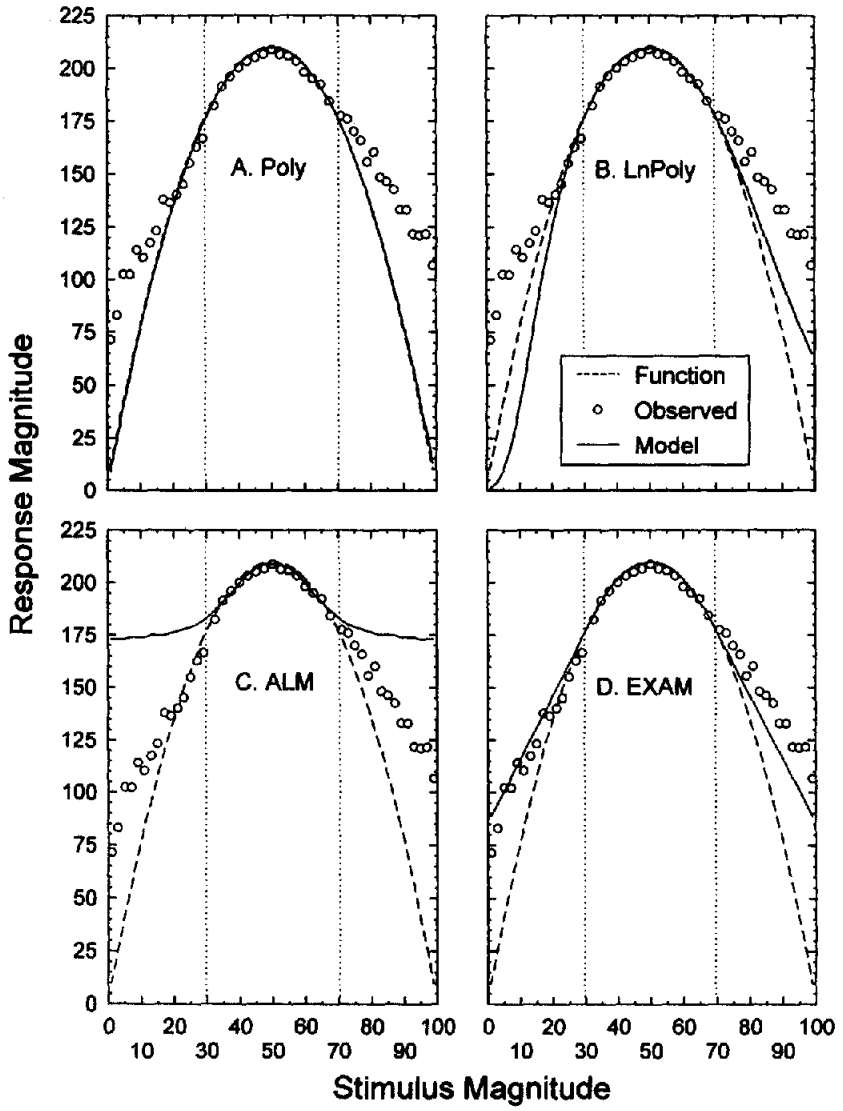

Figure 10. Model predictions and observed data for the quadratic function. Poly = polynomial hypothesis-testing model; LnPoly = log-polynomial adaptive-regression model; ALM = associative-learning model; EXAM = extrapolation–association model.

steps; therefore, we used $M = 201$ input nodes, $\{X_0 = 0, X_1 = 0.5, X_2 = 1, X_3 = 1.5, \ldots, X_{200} = 100\}$, with one input node corresponding to each possible stimulus magnitude. Likewise, each output node Y_j corresponded to a position on the real number line that was proportional to one of the possible response magnitudes. The outputs in the current experiments ranged from 0 to 250 units in single-unit steps; therefore, we used 251 output nodes, $\{Y_0 = 0, Y_1 = 1, Y_2 = 2, \ldots, Y_{250} = 250\}$, with one output node corresponding to each possible response magnitude. Note that these nodes covered the entire range of stimulus and response magnitudes, thereby allowing new interpolation or extrapolation responses to be produced on transfer tests.

When a particular stimulus X is presented, it activates the entire set of M input nodes, and each node is activated according to its similarity to the presented stimulus. The symbol $a_i(X)$ represents the activation of input node X_i when stimulus magnitude X is presented. A Gaussian activation function is assumed:

$$a_i(X) = \exp\left[-\gamma \cdot [X - X_i]^2\right], \tag{6}$$

where γ is a scaling parameter that determines the steepness of the generalization gradient.

Activation passes from the input nodes to output nodes as given by the following equation, where the activation of output node Y_j is denoted o_j, and the strength of association between each input node (X_i) and each output node (Y_j) is symbolized w_{ji}:

$$o_j(X) = \Sigma_{i=1,M} w_{ji} \cdot a_i(X). \tag{7}$$

The last term in this equation, the Gaussian activation function, yields response generalization. The probability that response Y_j is chosen from a set of L possible responses is given by the ratio rule:

$$P[Y_j|X] = o_j(X)/\Sigma_{k=1,L}o_k(X). \quad (8)$$

Thus, the response is chosen simply on the basis of the strength of its output activation. Finally, the mean output to stimulus X is the weighted average,

$$m(X) = \Sigma_{j=1,L}Y_j \cdot P[Y_j|X]. \quad (9)$$

The mean output given by this equation is used to predict participants' mean response to stimulus X.

The connection weight w_{ji} that associates input node (X_i) to output node (Y_j) is learned as follows: During feedback, the feedback signal Z activates each output node Y_j according to the Gaussian similarity function,

$$f_j(Z) = \exp\{-\gamma \cdot [Z - Y_j]^2\}, \quad (10)$$

where $f_j(Z)$ is the activation of output node Y_j by the feedback signal Z. The connection weights are updated according to the delta learning rule:

$$w_{ji}(t + 1) = w_{ji}(t) + \alpha \cdot [f_j[Z(t)] - o_j[X(t)]] \cdot a_i[X(t)]. \quad (11)$$

Knapp and Anderson (1984) used a Hebbian learning rule, whereas Kruschke (1992) used a delta learning rule; we adopted the latter because it is better supported.

The associative-learning model has two parameters: the scaling parameter (γ) and the learning rate (α). These parameters were estimated separately for each function condition k by searching for the lowest scale and smallest learning rate that would produce an MAE_k less than 1.2. The three rows labeled "ALM" in Table 1 show the estimated parameters and *MAE* values of ALM for the three function conditions. The scaling parameter and learning rate required to produce a sufficient fit (MAE_k less than 1.2) to asymptotic learning performance were smaller for the linear condition than for exponential and quadratic conditions. Note, however, that the best-fitting scaling parameters were too large (i.e., the generalization gradients were too narrow) to produce extrapolation. To examine the limits of extrapolation by ALM, we plotted the predictions by using a much smaller scaling parameter ($\gamma = .03$), which yielded a generalization gradient that covered the entire response continuum (but also produced an *MAE* that exceeded the data).

Panels C of Figures 8, 9, and 10 show the predictions of ALM for each function condition across the 45 transfer trials, using the small scaling parameter (instead of the parameter estimated from the training data). ALM roughly reproduced the assigned function in the interpolation region and also approximated the responses of participants in this region. In addition, ALM generated responses outside the range of trained responses when given extrapolation trials. Extrapolation was quite limited, however, and did not approach the extensive extrapolation observed in participants. We also examined model predictions across a wide range of learning rates and scaling parameter values. The fit of the model was insensitive to changes in the learning rate for $\alpha > 0.5$, and no value of the scaling parameter from $\gamma = 0.003$ to $\gamma = 3.0$ yielded substantial extrapolation. By failing to generate extreme extrapolation responses, ALM could not account for the data observed in the present study. This finding was not surprising given that ALM is based on ALCOVE (Kruschke, 1992), which was originally designed for category learning rather than function learning.

Extrapolation–Association Model (EXAM)

EXAM learns according to the same process assumed in ALM, but the response is constructed from a rule-based mechanism in accordance with Waganaar and Sagaria's (1975) observation that extrapolation is approximately linear. EXAM assumes that each input node corresponds to one of the training stimulus magnitudes and that each response node corresponds to one of the training response magnitudes. For the low-density training condition, for instance, there were eight stimulus–response pairs presented during training; therefore, $M = 8$ input nodes and $L = 8$ output nodes. Likewise, 20 and 50 input and output nodes were used for the medium- and high-density conditions, respectively. This assumption of EXAM is identical to that used in the association learning exemplar (ALEX) model proposed by Nosofsky and Kruschke (1992).[7]

The activation of input nodes and output nodes follows the same assumptions used in ALM, formalized in Equations 6 and 7. In addition, learning proceeds according to the same delta learning algorithm used by ALM (see Equation 11). The primary difference between EXAM and ALM is the mechanism used to generate responses. The first step in the response process is to match a presented stimulus to an input node corresponding to one of the training stimuli. The probability of matching stimulus X to input node X_i is

$$P[X_i|X] = a_i(X)/\Sigma_{k=1,M}a_k(X). \quad (12)$$

[7] Busemeyer, Byun, et al. (in press) developed a generalized version of EXAM that uses the same input–output coding assumptions as ALM (i.e., the same number of nodes as ALM). This generalized version of EXAM yields the same pattern of results as the version of EXAM presented herein. However, the generalized version of EXAM requires additional assumptions about the retrieval of previously trained stimuli, which complicates the model. We chose to present the simpler model here. Of course, it is also possible to use an alternate coding assumption with the ALM model so that it includes the same number of nodes as training values. This alternate version of ALM differs from EXAM only in terms of the mechanism for generating the response; therefore, the predictions of this alternative model can be obtained by using Equation 14 without the second term (which is responsible for linear extrapolation). However, this alternate version of ALM does not produce more extensive extrapolation than the original ALM model; thus, it is not considered in the present study.

Given that X is matched to X_i, three output values are retrieved: Output $Y(X_i)$ is retrieved by using a lower cue value X_{i-1} as the cue; output $Y(X_{i+1})$ is retrieved by using a higher cue value X_{i+1} as the cue; and output $Y(X_i)$ is retrieved by using the matching cue value X_i as the cue. The probability of retrieving each of the three outputs is given by Equation 8.

The next step is to select the response. Unlike ALM, which simply retrieves the strongest activated output, responses are generated using a cross-dimensional matching rule. The magnitude of the response is selected so that the proportion of change in output magnitudes matches the proportion of change in input magnitudes. For example, if the distance between transfer stimulus X and the matched cue X_i is only half of the total distance between the upper, X_{i+1}, and lower, X_{i-1}, cues, then the response is selected such that the distance between the response output Y and the matched-cue output $Y(X_i)$ is half of the total distance between the upper-cue output, $Y(X_{i+1})$, and the lower-cue output, $Y(X_i)$. This cross-modality matching process is mathematically formalized as follows:

$$[Y - Y(X_i)]/[Y(X_{i+1}) - Y(X_{i-1})] = [X - X_i]/[X_{i+1} - X_{i-1}],$$

which is algebraically equivalent to

$$Y = Y(X_i) + \{[Y(X_{i+1}) - Y(X_{i-1})]/[X_{i+1} - X_{i-1}]\} \cdot [X - X_i]. \quad (13)$$

Note that the response generated by Equation 13 is based on two parts. The first component is the retrieved output value, which is the response value associated with the training stimulus most similar to the transfer cue. This is the component on which ALM is based. The second, new component is responsible for linear interpolation and extrapolation using a slope value computed from retrieved instances.

On the basis of these assumptions, the mean response to transfer stimulus X is given by

$$E[Y|X] = \Sigma_{i=1,M} Pr[X_i|X] \cdot E[Y|X_i]$$

$$E[Y|X_i] = m(X_i) + [m(X_{i+1}) - m(X_{i-1})]/[X_{i+1} - X_{i-1}] \cdot [X - X_i], \quad (14)$$

where $m(X)$ is defined by Equation 9. If X_1 (the smallest input node) is most activated, then input node X_{i-1} in Equations 13 and 14 is replaced with input node X_i. If X_M (the largest input node) is most activated, then input node X_{i+1} is replaced with input node X_i. The mean rule-constructed output $E[Y|X]$ is used to predict participants' mean responses to stimulus X.

As with the associative-learning model, EXAM has two parameters to be estimated: the scaling parameter (γ) and the learning rate (α). We estimated these parameters separately for each function condition k by searching for the lowest scale and the smallest learning rate that would yield a MAE_k less than 1.2. The three rows labeled *EXAM* in Table 1 show the results of fitting this model to the learning data. The only significant difference between EXAM and ALM during the learning phase was the number of input and output nodes used to form associations; therefore, these two models made essentially the same learning predictions.

Panels D of Figures 8, 9, and 10 display the predictions of EXAM for each function condition across the 45 transfer trials, using the parameters estimated from the training data. EXAM accurately approximated the assigned functions in the interpolation region but extrapolated in a manner that did not reproduce the assigned functions in the extrapolation regions, as was observed empirically. Most important, EXAM's extrapolations corresponded very well with the pattern of over- and underestimation produced by participants (although EXAM performed less well for the lower extrapolation region of the linear function). Although it is not shown graphically, it is also noteworthy that EXAM produced very similar extrapolation patterns across the range of density conditions examined herein, as was found empirically.

We also examined the predictions of EXAM by using a common scale value (3) and learning rate (0.9) for the three function conditions. In this case, the model produced the appropriate over- and underestimation predictions for the exponential and quadratic functions but yielded a straight line for the linear function, contrary to participants' pattern of underestimation.

EXAM produced underestimation of the linear function when small learning rates and scale values were used for the following reason: Intermediate training stimuli (i.e., training stimuli that do not lie at the boundaries of the training domain) are learned more quickly than stimuli at the end points of the training domain because of generalized feedback from adjacent stimuli on both sides of the intermediate stimuli. When the learning rate is low and the generalization gradient is wide, the end stimuli are not learned as well as intermediate stimuli at the conclusion of training. The result of learning in EXAM is that responses gradually move from zero (the initial values) toward the feedback value. Thus, if the stimuli at the boundaries of the training domain are learned less well than intermediate stimuli, their associated responses will be underestimated. This pulls the slope down at the boundaries of the training domain, and because extrapolation in EXAM is based primarily on the slopes at the endpoints, the extrapolation rule underestimates the programmed function. This underestimation can be eliminated by increasing the learning rate, such that responses are highly accurate at the endpoints by the end of training.

In sum, of the four models we considered, EXAM was most accurate in accounting for the set of results obtained in the present study. By combining associative learning with rule-based responding, EXAM better explained the extrapolation of function-based concepts than either a pure associative-learning model or a pure rule-learning model.

General Discussion

Many natural concepts are best described as and represented by functions. The association between population magnitude and amount of pollution, for instance, is naturally thought of in terms of a function. One salient and important characteristic of a function-based concept is that it can be used to generate appropriate new responses in the presence of novel stimuli. If the population were to suddenly increase, for example, the function could be used to anticipate the amount and direction of change in pollution. Despite the predictive value of extrapolation, very little systematic research has explored extrapolation behavior in human learners or has examined the processes by which extrapolation occurs. This study represents the most comprehensive empirical and theoretical investigation of extrapolation with function-based concepts to date.

Empirical Findings

An important empirical issue concerning extrapolation is the extent to which humans are willing to extrapolate following restricted exposure to a small number of functionally related input–output pairs. The data reported herein show that learners extrapolated much beyond the range of learned responses, and they did so in the direction of the assigned function. This extrapolation behavior was observed for different types of functions (linear, exponential, and quadratic), training conditions (stimulus sets of 8, 20, and 50 unique stimuli), task instructions, and stimulus–response interpretations.

Although participants generally extrapolated in the direction defined by the assigned functions, systematic deviations were observed, and these deviations varied depending on the function form. In the linear condition, participants underestimated the assigned function in both extrapolation regions. This pattern has been replicated in a subsequent study using a negative linear function (DeLosh, 1995). In the quadratic condition, participants overestimated the assigned function in both extrapolation regions, and this pattern has been replicated as well (Byun, 1995; DeLosh, 1995). For the exponential function, participants also overestimated the function in both extrapolation regions.

The fact that systematic over- and underestimation were obtained only on extrapolation tests and not during training or on interpolation tests underscores the importance of examining extrapolation behavior. Moreover, these deviations indicate that extrapolation is more complex than inducing and applying a global rule. The underestimation obtained with the linear function condition was particularly telling. In this case, a simple linear rule was available and appropriate, but apparently many participants did not abstract or apply this rule. Had they done so, their transfer responses would have followed a straight line through the interpolation and extrapolation regions, but in fact, the observed curves for the linear training function were consistently nonlinear.

A second empirical issue concerns the number of unique stimulus inputs presented during training (i.e., density). From the perspective of verbal learning paradigms, the density manipulation was analogous to a manipulation of list length. On the basis of findings in the verbal learning domain, we would expect that the low-density condition would yield faster and more accurate learning than the higher density conditions (cf. Gillund & Shiffrin, 1984; Murdock, 1962; Roberts, 1972; Waugh, 1972). We were somewhat surprised, however, to find that density did not affect the rate of learning. One possibility is that the greater similarity of items in higher density conditions counteracted the disadvantages produced by longer lists (because list length was confounded with similarity in our density manipulation). Another possibility is that the higher density conditions were more likely to reveal the systematic relation among stimulus–response pairs (cf. Engelkamp, Biegelmann, & McDaniel, in press; Hunt & Seta, 1984), thereby facilitating learning. By this account, the disadvantage of learning more items in a longer list may have been balanced by the advantage of the salience in the relationship between items. Yet density did not affect either interpolation or extrapolation performance, a result that urges caution in accepting the idea that the higher density conditions facilitated abstraction of the stimulus–response relationship.

A third empirical issue concerns individual differences in what learners acquire when exposed to stimulus–response pairs generated from a function. On training and interpolation test trials, response accuracy was relatively high and individual differences were not apparent. Examining just these data might have led one to believe that all learners acquire the function very accurately. On extrapolation test trials, however, substantial differences among individuals were revealed. Some learners almost perfectly reproduced the training function across extrapolation tests, yet two others showed virtually no extrapolation outside the range of trained responses. Once again, this finding underscores the importance of using extrapolation tests to obtain a complete picture of what is acquired during function learning.

Theoretical Findings

In our theoretical consideration of the extrapolation results, we first evaluated three extant models of conceptual behavior: two prominent rule-based accounts of function learning and a prominent associative account of category learning. These models contrast two venerable theoretical approaches to concept learning. The associative-learning approach is exemplified by ALCOVE (Kruschke, 1992). Although ALCOVE was originally designed for category learning tasks, its success in the category-learning domain compels testing its usefulness in the function-learning domain. Moreover, ALCOVE includes stimulus generalization and can be modified to include response generalization as well (as is most appropriate when both stimuli and responses vary on continuous dimensions). With both stimulus and response generalization, the model can produce a limited amount of extrapolation. This extension of ALCOVE for function learning has been designated ALM.

Our application of ALM to function learning yielded two related findings. First, the extrapolation capabilities of ALM were very limited across the wide range of generalization gradients that we examined. Second, because ALM can extrapolate very little beyond the stimulus domain with which it is trained, it was unable to account for the extensive extrapolation produced by humans. Thus, associative models of the type evaluated herein do not provide an adequate account of extrapolation in function learning. This finding is not surprising, given that the class of associative models we tested was originally developed to explain category learning. Nonetheless, this finding is important because it underscores a salient limitation of nonabstractionist learning models and urges the incorporation of rule- or abstraction-based mechanisms in order to account for the entire range of human conceptual behavior (see Anderson & Fincham, 1996, for a related point).

The two rule-learning models that we evaluated were specifically designed to account for performance in function-learning tasks, and these models support extensive extrapolation. The polynomial hypothesis-testing model (Brehmer, 1974; Carroll, 1963) yielded extrapolation performance that perfectly matched (linear and quadratic conditions) or nearly perfectly matched (exponential condition) the assigned functions. Accordingly, this model failed to capture the pattern of extrapolation behavior of participants in this study, who systematically over- and underestimated the assigned functions when generating their extrapolation responses. It is interesting that the log-polynomial adaptive-regression model (Koh & Meyer, 1991) does produce extrapolation responses that deviate from the assigned functions. These deviations, however, were not uniformly consistent with those produced by human learners. For instance, in the high-extrapolation region of the linear function, the model overestimated the linear function, whereas participants underestimated the function; in the low-extrapolation region of the quadratic function, the model underestimated the function, whereas participants overestimated the function. Therefore, existing rule-based models of function learning do not reproduce the observed pattern of extrapolation performance.

Because neither associative-learning nor rule-based models of function learning successfully accounted for the empirical results observed in the present study, it appears that a new theoretical approach is necessary for understanding the processes by which humans learn and apply function-based concepts. A major contribution of this article is the development and testing of a new model of function learning (EXAM) that combines the associative learning of stimulus–response pairs with a response generation process that is based on linear interpolation and extrapolation. This hybrid model produced extrapolations much beyond the range of learned responses. Moreover, the extrapolations deviated from the assigned functions in a manner that corresponded to the pattern observed in human learners.

Recall that the response mechanism used by EXAM can be broken down into two parts (see Equation 13). One component is simply the retrieved output (similar to that used in ALM), and the second component produces linear extrapolation from retrieved outputs. Individual differences can be captured in this model by including the additional assumption that the second component is applied probabilistically, such that the linear-extrapolation rule is not used on every trial. Each participant would then have some probability of applying the second (linear extrapolation) part of the rule. A participant who consistently applies both parts will produce substantial extrapolation of a linear form, but a participant who only occasionally applies the second part will, on average, extrapolate to a lesser extent. In this manner, EXAM can account for the large individual differences observed in the present study.[8]

As a final point, it is important to note that although EXAM is currently limited in its applicability to function-learning tasks, we propose that the basic idea captured by the model represents a viable new approach to generalizing associative models of concept learning. The EXAM model is not simply a mixture of associative and rule-learning models, in the sense that EXAM does not generate extrapolation responses separately through associative and rule-learning mechanisms and then average the values or select the best value of the two for simulating human performance. Rather, the model incorporates (a) an associative-learning process like that used in connectionist implementations of exemplar-based models of categorization and (b) a response process that is guided by a rule operating on retrieved associations. As such, EXAM is an instantiation of our general view that a synthesis of associative and rule-based mechanisms is necessary for any general theory of conceptual behavior.

[8] The current version of the model assumes that information about the slope of a function at a training point is determined at the time of output retrieval. One could alternatively propose a model that learns and stores relational information (local slope values) during training. This model yields predictions that are essentially identical to those produced by EXAM. Note that a slope-learning model can account for individual differences if it assumed that there are different learning parameters for slope values and output values and that the learning rate parameter for slopes can vary across individuals or groups, depending on the extent to which they rely on relational information. For instance, a positive linear bias in the weight structure of the model combined with a slow learning rate for slope values yields no extrapolation in the upper region, similar to the pattern shown in Figure 5. Because our data did not delineate between slope-retrieval and slope-learning approaches, we adopted the simpler model.

References

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

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 259–277.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Bourne, L. E., Jr. (1966). *Human conceptual behavior.* Boston: Allyn & Bacon.

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

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (in press). Function learning based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories.* London: UCL Press.

Busemeyer, J. R., McDaniel, M. A., & Byun, E. (1997). The abstraction of intervening concepts 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 relation on function learning.* Unpublished doctoral dissertation, Purdue University.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional maps relating stimulus and response continua* (ETS RB 63-26). Princeton, NJ: Educational Testing Service.

Deane, D. H., Hammond, K. R., & Summers, D. A. (1972). Acquisition and application of knowledge in complex inference tasks. *Journal of Experimental Psychology, 92,* 20–26.

DeLosh, E. L. (1995). *Hypothesis testing in the learning of functional concepts.* Unpublished master's thesis, Purdue University, West Lafayette, IN.

Engelkamp, J., Biegelmann, U., & McDaniel, M. A. (in press). Relational and item-specific information: Trade-off and redundancy. *Memory.*

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W. K. (1995). *Classification and cognition.* New York: Oxford University Press.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–65.

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

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

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

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 454–464.

Kellogg, R. T., & Bourne, L. E., Jr. (1989). Nonanalytic–automatic abstraction of concepts. In J. B. Sidowski (Ed.), *Conditioning, cognition, and methodology: Contemporary issues in experimental psychology* (pp. 89–111). Lanham, MD: University Press of America.

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 616–637.

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

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

Ljung, L., & Soderstrom, T. (1983). *Theory and practice of recursive estimation.* Cambridge, MA: MIT Press.

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

Murdock, B. B., Jr. (1962). The serial position effect in free recall. *Journal of Experimental Psychology, 64,* 482–488.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

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

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, pp. 149–167). Hillsdale, NJ: Erlbaum.

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.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of the total-time hypothesis. *Journal of Experimental Psychology, 92,* 365–372.

Rumelhart, D. E., & McClelland, J. L. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Summers, S. A., Summers, R. A., & Karkau, V. T. (1969). Judgments based on different functional relationships between interacting cues and criterion. *American Journal of Psychology, 82,* 203–211.

Surber, C. F. (1987). Formal representation of qualitative and quantitative reversible operations. In J. Bisanz, C. J. Brainerd, & R. Kail (Eds.), *Formal operations in developmental psychology: Progress in cognitive development research* (pp. 115–154). New York: Springer-Verlag.

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

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

Waugh, N. C. (1972). Retention as an active process. *Journal of Verbal Learning and Verbal Behavior, 11,* 129–140.

(Appendixes follow)

Appendix A

Training Stimulus Magnitudes for the Low-, Medium-, and High-Density Conditions

Density condition	Training stimuli
Low	30.5, 36.0, 41.0, 46.5, 53.5, 59.0, 64.0, 69.5
Medium	30.0, 31.5, 33.0, 34.5, 36.5, 38.5, 41.0, 43.5, 46.0, 48.5, 51.5, 54.0, 56.5, 59.0, 61.5, 63.5, 65.5, 67.0, 68.5, 70.0
High	30.0, 30.5, 31.0, 32.0, 33.0, 33.5, 34.5, 35.5, 36.5, 37.0, 38.0, 38.5, 39.5, 40.5, 41.5, 42.0, 43.0, 43.5, 44.5, 45.5, 46.5, 47.0, 48.0, 48.5, 49.0, 51.0, 51.5, 52.0, 53.0, 53.5, 54.5, 55.5, 56.5, 57.0, 58.0, 58.5, 59.5, 60.5, 61.5, 62.0, 63.0, 63.5, 64.5, 65.5, 66.5, 67.0, 68.0, 69.0, 69.5, 70.0

Appendix B

Transfer Stimulus Magnitudes for Experiment 1

Transfer region	Transfer stimuli
Interpolation	32.5, 35.0, 37.5, 40.0, 42.5, 45.0, 47.5, 50.0, 52.5, 55.0, 57.5, 60.0, 62.5, 65.0, 67.5
Low extrapolation	1.0, 3.0, 5.0, 7.0, 9.0, 11.0, 13.0, 15.0, 17.0, 19.0, 21.0, 23.0, 25.0, 27.0, 29.0
High extrapolation	71.0, 73.0, 75.0, 77.0, 79.0, 81.0, 83.0, 85.0, 87.0, 89.0, 91.0, 93.0, 95.0, 97.0, 99.0

Received September 19, 1995
Revision received December 16, 1996
Accepted December 16, 1996 ■