Psychological Review
2004, Vol. 111, No. 4, 1072–1099

0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.4.1072

Population of Linear Experts: Knowledge Partitioning and Function Learning

Michael L. Kalish
University of Louisiana at Lafayette

Stephan Lewandowsky
University of Western Australia

John K. Kruschke
Indiana University

Knowledge partitioning is a theoretical construct holding that knowledge is not always integrated and homogeneous but may be separated into independent parcels containing mutually contradictory information. Knowledge partitioning has been observed in research on expertise, categorization, and function learning. This article presents a theory of function learning (the population of linear experts model—POLE) that assumes people partition their knowledge whenever they are presented with a complex task. The authors show that POLE is a general model of function learning that accommodates both benchmark results and recent data on knowledge partitioning. POLE also makes the counterintuitive prediction that a person's distribution of responses to repeated test stimuli should be multimodal. The authors report 3 experiments that support this prediction.

The learning of concepts by induction from examples is fundamental to cognition and ". . .basic to all of our intellectual activities" (Estes, 1994, p. 4). Many concepts are categorical: for example, when a paleontologist learns to classify dinosaurs as bird-hipped or lizard-hipped, when an infant learns to label furry four-legged animals as cats or dogs, or when a physician learns to categorize a nevus as benign or potentially cancerous. In these cases, responses are limited to a nominal scale, often consisting of binary response options such as "Category A" or "Category B."

However, people often also learn function concepts, in which a continuous stimulus variable is associated with a continuous response variable. For example, one may learn how long to water the lawn as a function of the day's temperature, how driving speed affects stopping distance, what his or her blood alcohol level will be depending on the number of cocktails consumed, and so on. Function concepts thus subsume category concepts as the small subset of cases in which the response scale is nominal rather than continuous. Remarkably, cognitive psychology to date has devoted far more empirical and theoretical attention to categorization than to function concepts as a whole.

The purpose of this article is twofold. First, we seek to raise the profile of function concepts by presenting a computational theory of function learning that is based on the idea that people simplify a complex learning task by partitioning it into multiple independent modules. The theory, known as POLE—for population of linear experts—is shown to handle most existing data on function learning. Three new experiments explore some of POLE's counterintuitive predictions and provide additional support for the theory. We show that when people are confronted with uncertainty about which of several competing functions applies to a test stimulus, responding alternates between different learned functions rather than relying on a blend of existing knowledge, thus giving rise to multimodal response distributions.

The second purpose of this article is to evaluate an overarching framework for learning and knowledge acquisition, known as *knowledge partitioning* (e.g., Lewandowsky, Kalish, & Ngang, 2002; Lewandowsky & Kirsner, 2000; Yang & Lewandowsky, 2003, in press). Knowledge partitioning holds that people's knowledge is often heterogeneous and divided into independent parcels that may contain mutually contradictory information. Knowledge partitioning has been identified in experts (Lewandowsky & Kirsner, 2000) as well as in nonexperts in category learning (Lewandowsky, Kalish, & Griffiths, 2000; Yang & Lewandowsky, 2003, in press) and function concept acquisition (Lewandowsky et al., 2002). Here, we show that knowledge partitioning not only is a phenomenon in its own right but also is fundamental to function learning.

We proceed as follows: We first provide a brief overview of function learning before turning to a discussion of knowledge partitioning and its differentiation from empirical precursors. This leads us to present the new theory of function learning, POLE. The

Michael L. Kalish, Institute of Cognitive Science, University of Louisiana at Lafayette; Stephan Lewandowsky, School of Psychology, University of Western Australia, Crawely, Western Australia, Australia; John K. Kruschke, Department of Psychology, Indiana University.

Preparation of this article was facilitated by a Large Research Grant from the Australian Research Council to Michael L. Kalish and Stephan Lewandowsky and a Linkage International Grant from the Australian Research Council to Michael L. Kalish, Stephan Lewandowsky, and John K. Kruschke. We thank Leo Roberts for his assistance during data collection and manuscript preparation. Thanks to Barbara Dosher, Gordon Logan, Simon Farrell, and Matthew Duncan for their helpful comments on a draft. For additional assistance with data collection, we thank Melissa Adkins, Nancy Aleman, Twanna Allen, Dan Hall, and Rob Koleszar.

Correspondence concerning this article should be addressed to Michael L. Kalish, Institute of Cognitive Science, University of Louisiana at Lafayette, Lafayette, LA 70504-3772. E-mail: kalish@louisiana.edu

fundamental assumption of POLE is that people partition their knowledge into numerous independent components; we show that the model can quantitatively account not only for the partitioning phenomena that stimulated its development but also for all benchmark results in the function-learning literature. This identifies knowledge partitioning as being fundamental to function concept learning, which extends a similar conclusion reached in the context of category learning by Yang and Lewandowsky (in press; see also Erickson & Kruschke, 1998). We conclude the article with three experiments that provide strong support for one of POLE's most counterintuitive predictions, namely, that people in situations of uncertainty alternate between different learned functions.

LEARNING FUNCTION CONCEPTS

In the domains that are commonly studied by cognitive psychologists, functions play at least a supporting role in much of decision making. For example, on a recent trip by the authors to the Black Forest in Germany, the water main broke in our street. The experienced repair worker arrived with a large wrench, one end of which he placed on the broken pipe's valve, through which the leaking water was flowing. The other end he put to his ear, explaining that he could tell the distance to the break from the frequency of the sound. True to his word, his excavation of the street was within 1 m of the break, over 6 m from his listening post. His estimate of the function relating frequency to distance was learned over many years' experience in the field. In the laboratory, the nature of continuous judgments of this type is best examined with a function-learning methodology.

In a function-concept learning paradigm, people learn the relationship between continuous stimulus and response dimensions from a set of discrete training items. On each learning trial, the magnitude of the stimulus dimension is presented, and the participant's task is to predict the associated response magnitude. Each response is followed by corrective feedback. Typically, stimulus magnitudes are coded graphically and without any explicit numeric coding, for example, by a horizontal arrow of varying lengths (Lewandowsky et al., 2002). Response magnitudes can be variously provided by response keys that are labeled with discrete numbers (Kruschke, 2001a), written or verbal responses (Birnbaum, 1976; Mellers, 1986), typing of numeric values (Reed & Evans, 1987), the elapsed time between two key presses (Koh & Meyer, 1991), and graphical means (e.g., participants might have to adjust a vertical slide rule with the mouse, with corrective feedback being presented on the same slide rule; Lewandowsky et al., 2002).

The learning of function concepts shares a certain similarity with the learning of sensory and motor functions (Rosenbaum, Carlson, & Gilmore, 2001). In prism adaptation experiments, for example, participants might be presented with a visual target and asked to touch it or throw an object at it, thus learning to compensate for the distortion by prism goggles (Martin, Keating, Goodkin, & Bastian, 1996; Welch, Bridgeman, Anand, & Browman, 1993). Dynamic touch (Turvey, 1996), vestibulo-ocular reflex gain adaptations (Crawford & Guitton, 1997), depth cue integration (Atkins, Fiser, & Jacobs, 2001), motor field adaptations (Conditt, Gandolfo, & Mussa-Ivaldi, 1997), and adaptation to visuomotor distortions (Ghahramani & Wolpert, 1997) all appear to involve rapid learning of continuous input–output mappings. However, all these cases differ from function learning in two important ways. First, whereas function concepts involve highly abstract knowledge that can be tested and used in a variety of response modes, most sensory-motor functions are highly response specific (e.g., an acquired adaptation to a visual distortion cannot be reported verbally or by pressing keys). Second, all research on sensory-motor adaptations involves preexisting functions, based on life-long practice, that may be modulated during an experiment but are not learned de novo. Here, we focus on abstract function-concept learning and only briefly point to possible connections between function concepts and motor learning later in the article.

With sufficient practice, people are remarkably adept at learning a variety of function concepts (e.g., Busemeyer, Byun, DeLosh, & McDaniel, 1997). The discrete stimuli presented during training are generalized into a continuous function embodying the underlying relationship. This is evident during a transfer phase when novel stimuli are presented and participants must produce the associated response value. If those novel stimuli fall within the range of training values, performance is typically highly accurate (see Busemeyer et al., 1997, for a review). If novel stimuli are presented outside the range of training values, participants are capable of extrapolation, albeit at a lower level of accuracy than interpolation (DeLosh, Busemeyer, & McDaniel, 1997). Although it is possible for people to learn functions that map multiple stimulus dimensions onto a single response dimension (e.g., Koh, 1993; Roe, Barkan, & Busemeyer, 2001), most research to date has been conducted with one-dimensional functions in which a single stimulus property determines each response. This article is primarily concerned with one-dimensional function learning.

Because the human ability to use concepts of any type rests on the kind and structure of knowledge that is acquired during training, we consider additional related areas of research in the context of our second goal—demonstrating that contrary to much consensus, knowledge is not always integrated and homogeneous.

HOMOGENEITY VERSUS HETEROGENEITY OF KNOWLEDGE

There are many and varied theories of knowledge acquisition and representation. Notwithstanding their theoretical diversity, we suggest that at least one common theme can be identified: On balance, current views of knowledge tend to emphasize integration and homogeneity, rather than leaning toward contradiction, idiosyncrasy, and heterogeneity. Lewandowsky et al. (2002) explored this theme in some detail, noting in particular the common tenet that expert knowledge is highly organized and integrated (e.g., Bédard & Chi, 1992; Ericsson, 1996; Glaser, 1996) and the nearly uniform view among theories of categorization that all available knowledge enters into classification decisions (either because all disparate experiences have been combined into a single integrated rule or prototype—Ashby & Gott, 1988; Homa, Sterling, & Tepel, 1981—or because all remembered instances are considered during classification—Kruschke, 1992; Nosofsky, 1991).

Lewandowsky et al. (2002) additionally identified a pervasive boundary to this integration theme: Most theorists agree that knowledge is specific to a domain or situation and that it may become inaccessible through context shifts or when a domain-relevant problem is made atypical. Hence, chess masters recall randomly arranged chess boards with great difficulty, relative to midgame positions (e.g., Gobet & Simon, 1996), and people consistently fail to apply a known solution strategy if a problem

isomorph is presented in a novel context (e.g., Gick & Holyoak, 1980; Holyoak & Koh, 1987). Likewise, the integration assumption acknowledges the coexistence of several complementary rules or strategies that may tap the same knowledge (e.g., Erickson & Kruschke, 1998; Lovett & Schunn, 1999; Shrager & Siegler, 1998). In the perceptual domain, people are able to learn multiple (correct) visuomotor mappings, cued, for example, by the presence or absence of either a tone (Kravitz & Yaffe, 1972) or prisms (Martin et al., 1996).

However, the integration assumption does entail the expectation that alternative strategies cooperate rather than compete, irrespective of the context in which performance is observed. Accordingly, the pervasive, if often tacit, expectation of much current theorizing is that although performance on a given problem may differ between contexts, responses should nonetheless remain relatively free of contradiction.

Violations of Homogeneity: Expertise

There are, however, strong reasons to question the integrality and homogeneity of knowledge. We consider evidence from the domains of expertise, categorization, and function learning in turn.

Given the emphasis on knowledge integration within the expertise literature, the occurrence of contradictory behavior among experts is particularly striking (e.g., Carraher, Carraher, & Schliemann, 1985; Lewandowsky & Kirsner, 2000; Nunes, Schliemann, & Carraher, 1993; Schliemann & Carraher, 1993). For example, Lewandowsky and Kirsner (2000) asked experienced fire commanders to predict the spread of simulated wild fires from a set of physical predictors. When the two key predictors (wind and slope of terrain) were in opposition—thus making direction of spread uncertain—the experts' predictions depended on the physically irrelevant problem "context": namely, the origin of the fire. When fires were presented as to be controlled, experts uniformly expected them to spread with the wind, whereas physically identical fires presented as back burns (i.e., fires lit by fire fighters as a countermeasure ahead of the to-be-controlled fire) were expected to spread into wind. These contradictory responses ignore the fact that back burns obey the same laws of physics as any other fire.

To facilitate understanding of the remainder of this article, we must differentiate the finding of Lewandowsky and Kirsner (2000) from conventional context effects. This differentiation rests on the following criteria:

1. The nature of the problem and its surface structure did not differ between contexts.

2. The domain relevance of the problem was equal across contexts as both types of fire are routinely considered during training.

3. The change in context was a mere change of a verbal label accompanying the problem that did not alter the surface structure of the problem itself. (We consider only this local meaning of *context* from here on.)

4. The context shift did not merely impair performance but engendered a qualitative reversal of the response. That is, the same problem yielded two mutually exclusive and contradictory predictions, each of which was consistent with application of a domain-relevant predictor variable.

5. The test context was objectively irrelevant and did not, by itself, predict fire spread on the basis of any physical laws.

6. Most critically, a single unified functional relationship existed that predicted the problem outcome in both contexts and that the experts explicitly learned during their extensive training.

7. All predictor variables of this unified mechanism were equally valid in both contexts.

The preceding criteria clarify that earlier demonstrations of context specificity of knowledge did not represent partitioning: For example, the lack of transfer between isomorphs observed in problem solving (e.g., Holyoak & Koh, 1987) does not meet Criteria 1–4; the disruption of expert performance with randomized chess boards (e.g., Gobet & Simon, 1996) does not satisfy Criteria 2 and 3; context-specific gaze adaptations to different prisms (Welch et al., 1993) do not meet Criteria 1, 2, 3, and 5; and so on.

Lewandowsky and Kirsner (2000) explained their finding that experts would make two mutually opposing predictions for the same problem in two domain-relevant contexts by postulating the concept of knowledge partitioning. According to this framework, knowledge may be divided or modularized into independent "parcels," with the strong possibility that inconsistent or mutually exclusive information persistently coexists. Critically, those diverse knowledge components may be brought to bear on the same invariant problem.

There is, however, at least one potential alternative explanation for the results of Lewandowsky and Kirsner (2000). Fire experts in the field are statistically more likely to encounter wind-driven than slope-driven to-be-controlled fires (and vice versa for back burns). In consequence, notwithstanding its physical irrelevance, context is a psychologically valid probabilistic predictor of fire spread. Using a categorization analogue of the fire prediction task, Lewandowsky et al. (2000) showed that this probabilistic mapping between context and outcome may engender the appearance of contradiction without, however, mandating the assumption critical to knowledge partitioning: namely, that a nonpredictive context cue selectively gates access to independent and mutually contradictory knowledge parcels. We next turn to experimental support for the knowledge partitioning framework that has addressed this issue and that also returns us to the function concept arena.

Heterogeneity in Function Learning and Categorization

Lewandowsky et al. (2002, Experiments 1 and 2) implemented an analogue of fire prediction in a function-learning task, using a function that related speed of fire spread to (downhill) wind speed for a particular constant slope. The function was an upward concave quadratic and, although arbitrarily parameterized, captured the true physical situation. The vertex of the function represented the wind speed at which the direction of the fire reverses: Above that point, the force of wind is sufficient to overcome the effect of slope, thus blowing the fire downhill with increasing speed. Below that point, the wind is insufficient to blow the fire downhill, and speed of spread uphill increases with decreasing wind strengths.

Components of this common to-be-learned function were preferentially shown in two different contexts during training, by

labeling each stimulus as either a back burn or a to-be-controlled fire. In a randomized condition, training stimuli across the entire range of the function were presented in both contexts, whereas in a partitioned condition most back burns were associated with low wind speeds (and hence the decreasing component of the quadratic function) and most to-be-controlled fires with high wind speeds (and hence the increasing component). In both conditions, however, context was uncorrelated with the correct magnitude of the response variable, speed of fire spread.

At a subsequent transfer test, all stimuli were shown in both contexts. Responses revealed that participants in the randomized condition learned an integrated quadratic function, whereas participants in the partitioned condition, by contrast, appeared to learn the two segments of the function separately, as indicated by their context-specific extrapolations. These context-specific extrapolations, moreover, were highly correlated with the responses of people in two control conditions who learned only one segment of the quadratic function. The high correlations suggest that people in the partitioned condition had mastered each segment of the function in its preferential context as though the other had never been presented.

Yang and Lewandowsky (2003, in press) recently reported corresponding results in category learning: Context again identified component boundaries of the category space, without however being predictive of category membership itself. People's generalizations to novel transfer items were found to be context specific and, within each context, much like those of people who had learned only one or the other component boundary. This underscores the generality of the partitioning observed in function learning by Lewandowsky et al. (2002).

In interpreting the results of Lewandowsky et al. (2002; see also Yang & Lewandowsky, 2003, in press), we must reiterate some of the earlier factors that are particularly crucial to identifying knowledge partitioning: (a) Context, by itself, was not a direct predictor of the outcome; (b) a common context-invariant solution existed that people were demonstrably able to learn when context was randomly assigned to stimuli; and (c) all relevant variables were equally predictive in both contexts, and it was only their relationship to the outcome that changed with context. These factors, plus the additional fact that participants had no prior knowledge about the functions or categories, differentiate the results of Lewandowsky and colleagues from otherwise loosely related precedents in the domain of perceptual learning (e.g., Atkins et al., 2001; Ghahramani & Wolpert, 1997; Jacobs & Fine, 1999; Martin et al., 1996; Vetter & Wolpert, 2000).

The data of Lewandowsky et al. (2002; see also Yang & Lewandowsky, 2003, in press) are therefore arguably unique in suggesting that context can serve to gate a relevant parcel of knowledge that is learned independently of competing knowledge held elsewhere. We now take the idea of knowledge partitioning to its extreme, by postulating that partitioning is at the heart of all learning of function concepts.

FROM PARTS TO THE WHOLE: POLE, A NEW THEORY OF FUNCTION LEARNING

We propose that people, when faced with learning a complex prediction task, seek to break the problem down into simple constituent components and then learn those components independently and without attempts at cross-referencing or integration. We explore this view by presenting a new model of function learning, called POLE, which postulates that people respond by selecting the output of one of many possible component simple functions. No integration across those component functions is assumed to take place. This core property differentiates POLE from existing theories of function learning.

Models of function learning have been variously based on pure parametric function estimation (Koh & Meyer, 1991), on pure instance-based learning (ALM; Busemeyer et al., 1997), and on a combination of linear regression and instance-based learning. This last model, EXAM (for *ex*trapolation–*a*ssociation *m*odel), has accounted for a wider range of data than either pure rule-based or pure instance-based alternatives (DeLosh et al., 1997). We therefore consider EXAM to be a standard against which all new theories must be evaluated. However, as developed to date by DeLosh et al. (1997), EXAM is not suitable for fitting the partitioning results of Lewandowsky et al. (2002) because it cannot handle the presence of a binary context variable without significant modifications.

Elements of a New Theory

We propose that each training stimulus is associated not directly with a response, as in EXAM, but with a function that predicts the target magnitude. We propose furthermore that stimuli presented in different contexts can be selectively associated with different functions. For example, participants in the Lewandowsky et al. (2002) paradigm might learn one function (fire speed $\propto$ wind speed) for one context and another function (fire speed $\propto 1 -$ wind speed) for the other context.

POLE assumes that people are predisposed to expect a new function to be a maximally simple positive linear function ($Y \propto X$). This predisposition is subject to modification by experience; people may learn to use a function other than the one initially preferred, or if necessary, they may break the concept into component functions based on the values of one or more dimensions of the stimulus. The assumption of POLE is that, in all cases, stimuli are associated with simple candidate response functions (which, for parsimony only, are assumed to be linear), only one of which is chosen to provide a response on any given trial. POLE thus rejects the idea that people blend, average, or otherwise integrate information held in different components of knowledge.

Recent models in machine learning (Schaal & Atkeson, 1998), motor learning (Haruno, Wolpert, & Kawato, 2001), and category learning (Erickson & Kruschke, 1998), all based on the mixture-of-experts approach (Jacobs, Jordan, Nowlan, & Hinton, 1991), share some properties with POLE; we explore these relations further in our General Discussion.

Population of Linear Experts (POLE)

POLE models the psychological processes that take a potentially multidimensional stimulus x and produce a one-dimensional numeric response; the model learns by comparing the value of its current response, $\hat{y}$, to the target value, y. We implement the model as a connectionist network, paralleling a related model of categorization, ATRIUM (Erickson & Kruschke 1998, 2002; Kruschke & Erickson, 1994), that introduced the mixture-of-experts formalism (Jacobs et al., 1991) to category learning. An important conceptual difference between POLE and ATRIUM is that POLE generates a

response by selecting just one expert on any given trial, whereas ATRIUM uses a weighted mixture of the predictions of all experts.

The basic structure of the network is presented in Figure 1. When a stimulus is presented, it activates a set of nodes ("Instance Nodes" in Figure 1) that represent memory of previously encountered stimuli. Instances are arrayed according to their magnitude and may optionally be separated along a second, nominal context dimension. Instances have overlapping receptive fields and are activated proportionally to their similarity to a presented stimulus. In addition to activating instances, a presented stimulus also elicits a set of potential responses from an entire population of preexisting linear response functions ("Experts" in Figure 1), each with a unique slope and intercept. Although each function computes a candidate response, only one is selected as the network's output.

The choice of function is determined by a gate placed on each expert's output, whose value depends on two factors: (a) the activation of the instance nodes and their learned associations with each expert (solid lines in Figure 1) and (b) the strength of the bias ("Biases" in Figure 1) associated with each function. Each instance is connected to each function's gate such that positive connections indicate an increased probability that the function will be chosen to govern the output of the network. Conversely, negative connections indicate an increased probability that the function's predictions will be ignored. The bias term associated with each function reflects the model's instance-independent global preference for that function.

Learning takes place on any trial on which feedback about the correct response is available. Learning is not restricted to the function that was selected for the response. Instead, all candidate functions are evaluated, and the connections from instances to the gates of those functions whose predictions were more accurate than the average prediction (weighted by their choice probabilities) are strengthened, whereas connections to functions that made worse-than-average predictions are weakened. Biases for each function are also adjusted during this step. Once changes in choice probability are determined, the model then adjusts attention to the stimulus dimensions so as to maximize activation of the instance nodes responsible for selecting the correct experts.

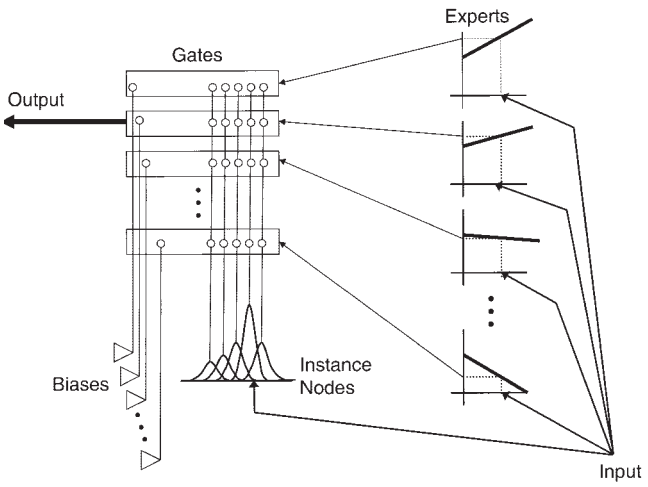

Figure 1. Architecture of POLE (population of linear experts model). Activation in the model flows from the inputs, via learned dimensional attention (not shown), to instance nodes. In parallel, each expert computes a possible response. Learned weights (open circles) connect all exemplars and expert-specific biases to gate nodes. The gates select one expert prediction to become the output, on the basis of the weighted activation from exemplars and biases.

The model thus embodies several hypotheses about the psychological processes responsible for function learning. Chief among these are dimensional attention, stimulus-specific knowledge, probabilistic selection of competing representations, and error-driven learning. We now present a more detailed description of each of these processes.

Dimensional Attention

Dimensional attention refers to the relative amount of processing given to each component of a stimulus. In categorization, attention can be readily conceptualized as the weighting of each dimension. If a dimension attracts attention, its values are more easily discriminated; if a dimension does not attract attention because it is irrelevant or redundant, values along that dimension become more difficult to discriminate. There is now much evidence to support the existence of dimensional attention and the long-term learning of such attention in category learning (Goldstone & Steyvers, 2001; Kruschke, 1992, 2001a, 2001b; Nosofsky, 1986). Accordingly, POLE also relies on the concept of dimensional attention.

In POLE, stimuli are represented by a vector $\mathbf{x}$, whose elements are restricted to the unit interval (0, 1). For the one-dimensional function-learning situations considered in this article, when a context-free stimulus is presented, $\mathbf{x}$ is one-dimensional, and its single dimension x_m represents numeric magnitude using an interval scale or better. In the one-dimensional case, there is no meaningful sense of dimensional attention; the magnitude dimension gets all attention devoted to the stimulus. In cases in which a stimulus is presented in an explicit context, as in the earlier experiments of Lewandowsky et al. (2002), $\mathbf{x}$ additionally includes a nominal dimension that codes context (represented here by 0 and 1 throughout). Dimensional attention then serves to weight the relative importance of context and numeric magnitude, thus translating the physical stimulus into a psychological representation (Nosofsky, 1986).

Each dimension i in $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is assigned an initial salience or baseline attraction, denoted $\aleph_i$. The common assumption that attention is of limited capacity implies that there is competition between dimensions, such that their attractions trade off against each other. This is implemented by normalizing the attractions to determine the final dimensional attention values, denoted α_i. Thus, the dimensional attention value for a given dimension I (we follow the convention that uppercase subscripts represent a fixed index for which a computation is completed and lowercase subscripts denote a varying index over which summation occurs) is given by

$$\alpha_I = \exp(\aleph_I) / \sum_i \exp(\aleph_i). \quad (1)$$

Note that α is necessarily equal to unity for the single dimension in $\mathbf{x}$ if one-dimensional stimuli are presented.

Instance Representations

Dimensional attention contributes to determining the activation of the instance nodes (a_j^{Inst}), whose overall profile of activation constitutes the representation of stimulus **x**:

$$a_j^{\text{Inst}} = \exp(-c \sum_i \alpha_i |x_i - \mu_{ji}|), \quad (2)$$

where c is a free parameter determining the specificity of the activation and μ_{ji} is the location of instance node j along dimension i. Locations of instance nodes are assumed to cover the entire stimulus space uniformly.

Conceptually, Equation 2 states that an instance is maximally activated by a stimulus of identical magnitude. Activation declines exponentially with an increase in the distance between the stimulus and the instance, where the steepness of that decline is determined by the specificity parameter c. The relative weighting of dimensions (e.g., context vs. magnitude) is a function of the distribution of attention as determined by Equation 1. Thus, a dimension that receives little attention will not cause activation to drop off as x and μ move away from each other, consistent with the basic assumption that ignored dimensions cannot influence discriminability of stimuli.

Population of Experts

The stimulus, represented by the activated instance nodes, provides a basis for choosing which of the possible linear response functions (experts) will be used on a given trial. Each of the k experts makes a prediction based on the simple function

$$\hat{y}_k = \beta_{0k} + \beta_{1k} x_m, \quad (3)$$

where x_m is the value of the magnitude dimension in **x** and the elements of β are the slope and intercept parameters.

These experts form the core of POLE's architecture. In the simulations reported below, there were 64 unique preexisting experts, each with a predetermined constant slope (β_1) and intercept (β_0). Slopes and intercepts were chosen subject to two constraints: First, for any magnitude of the stimulus x, there had to be a sufficiently large number of experts to cover the entire range of possible response magnitudes at a satisfactory resolution. Second, as many experts as possible were to be consistent with any given response magnitude for each stimulus value. These constraints were met by dividing the range of y into a number of intervals and mapping linear functions so that they intercepted the y-axis at $x = 0$ and $x = 1$ at these intervals. For example, if there were 64 experts, the ordinate was divided into eight intervals from -0.5 to 1.5 (the intervals exceed the range so that linear functions that cover only part of the function space can be used as experts). At each of these eight points along the ordinate (i.e., $x = 0$), eight functions were drawn whose $f(x)$ was equal to the same eight y-values when $x = 1$. This ensured that the 64 experts covered the entire function space by symmetrically fanning out from the same y-values at $x = 0$ and at $x = 1$.

Choice of Expert

The probability that any particular expert is chosen is determined jointly by two factors. First, each expert has some stimulus-independent probability of being chosen, reflected by a set of biases that embody a priori expectations combined with learned preferences. We chose to represent people's a priori expectations with an exponential gradient, centered on the experts with unit slope (i.e., $y = \beta_0 + x$) and decaying as the slope diverged from unity. This gives an initial bias for each expert:

$$w_{k0} = \omega \times \exp(-\varepsilon |M - m_k|), \quad (4)$$

where ω is the maximum initial bias and ε is the rate of decrease in bias as the slope of the expert (m_k) diverges from the preferred slope, M. On the basis of relevant data (Busemeyer et al., 1997), the preferred slope (M) was set to unity in all simulations below. Second, each instance node (μ_{ij}) is connected to each expert k by a weighted connection. The product of these two factors determines the strength of each expert. Similar to dimensional attention, final strength values are normalized. Before normalization, strength s_k of an expert is given by

$$s_k = w_{k0} \exp(\sum_j w_{kj} a_j^{\text{Inst}}), \quad (5)$$

where w_{k0} is the bias for expert k and w_{kj} is the weight to expert k from instance node j.

In POLE, the strength of an expert is interpreted as the probability of selecting that function's output as the response, and for this reason, strengths must always sum to unity, which is achieved by normalizing the s_k values. According to this constraint, the final strength S_K of an expert K is given by

$$P(K|\mathbf{x}) = S_K = \frac{s_K}{\sum_k s_k}. \quad (6)$$

Equation 6 also clarifies that the final normalized strength of an expert (S_K) is identical to the probability of that expert being chosen for the overt response when the stimulus **x** is presented. If an expert K is chosen to respond on a given trial, the output of the model, $\hat{y}$, will be $\hat{y}_K$ as defined by Equation 3.

The discrete choice of an expert on each trial implies that, across replications, POLE predicts a distribution of responses for a given stimulus. The nature of that distribution is determined by the output ($\hat{y}_K$) of each expert and the probability $P(K|\mathbf{x})$ of it being chosen. The ability to predict response variability, and the shape of the response distribution, is one of the crucial properties of POLE that we examine in detail later in three experiments. This discrete choice of expert on each trial most clearly distinguishes POLE from other models such as EXAM and ATRIUM, in which the output of the model is computed by averaging the potential responses of many instance nodes, weighted by the probability of each node's response being chosen (see Appendix A, here, and Equation 14 in DeLosh et al., 1997, and Equation 6 in Erickson & Kruschke, 1998).

In summary, the predictions of the model are determined by four factors: the attention to individual stimulus dimensions α_i, the bias toward each individual expert w_{k0}, the item-specific attention weights connecting instance nodes to experts (w_{kj}), and the specificity of instance nodes c. The first three factors are all adjusted dynamically to reduce error during learning, and the last factor is

represented as a free parameter.[1] The remaining free parameters of the model govern the error-driven learning processes.

Error Reduction During Learning

Although the output from only one expert is chosen on each trial, POLE learns by adjusting its entire response distribution to reduce error. In POLE, the error (E) for expert K is taken to be the squared difference between the target (y) and that expert's response ($\hat{y}_K$):

$$E_K = \tfrac{1}{2}(y - \hat{y}_k)^2. \quad (7)$$

The ½ in Equation 7 is a convenience to simplify the derivatives of the parameters with respect to error.

We then compute a strength-weighted average, or mixed, error (E_{Mix}) from all of the individual expert's errors:

$$E_{\text{Mix}} = \sum_k S_k E_k. \quad (8)$$

When a participant receives corrective feedback, expert strengths are adjusted to reduce this mixed error to as small a value as possible. In practice, because several experts may make similar predictions for any single stimulus, a number of experts may retain their strengths. As presented in Equation 6, final expert strength is normalized, and it is the prenormalized strengths that are adjusted to reduce error for a given expert K:

$$\Delta s_K = \eta_s \frac{(E_{\text{Mix}} - E_K)}{\sum_k s_k}, \quad (9)$$

where η_s is the shift rate (derivations of this and all other learning rules are presented in Appendix B). Because of the inherent nonlinearity in the relationship between these strengths and the average error, this shift is repeated 10 times per trial in the simulations reported below. With this in mind, we introduce a new notation for the strengths, so that s_k^{init} indicates the initial values given by Equation 5 and s_k^{shift} are the strengths at the end of the shifting process but still before normalization. Essentially the shift ensures that on each learning trial, experts that are producing smaller errors than average receive a boost in strength and experts that are making larger errors have their strengths reduced as a result of each shift. Iterating ensures a closer approximation to this gradient descent.

The shifted strength values are then learned, so that the same stimulus presented again will benefit from the shift and elicit a more accurate response distribution. The shifted strength values serve as targets for the initial strengths, so that POLE can learn by doing gradient descent on this difference for both the bias strengths (which have a stimulus-independent effect on choice probabilities)—

$$\Delta w_{k0} = \lambda_b (s_k^{\text{shift}} - s_k^{\text{init}}) \exp\Big(\sum_j w_{kj} a_j^{\text{Inst}}\Big) \quad (10)$$

—and for the weights associating specific instance nodes to the experts—

$$\Delta w_{kj} = \lambda_w (s_k^{\text{shift}} - s_k^{\text{init}}) s_k^{\text{init}} a_j^{\text{Inst}}, \quad (11)$$

where λ_b and λ_w are the bias and associative weight learning rates, respectively. The learning of biases enables the model to develop general preferences for certain functional relations, and the learning of weights enables the model to associate specific response functions with certain stimuli.

The final step on each trial is for the model to adjust dimensional attention (as opposed to the strengths of the experts, which was done using Equations 10 and 11). Conceptually, this is simply a case of descending the gradient of the dimensional attentions with respect to error; practically, the derivative is somewhat complex. For any dimension I, the change in (prenormalized) dimensional attractiveness is

$$\Delta \aleph_I = -\lambda_{\text{dim}} \sum_{k,j,i} (s_k^{\text{shift}} - s_k^{\text{init}}) s_k^{\text{init}} w_{kj} a_j^{\text{Inst}} |x_i - \mu_i| (\kappa_{iI} - \alpha_i) \alpha_I, \quad (12)$$

where κ_{iI} is the Kronecker delta; hence κ_{iI} equals 1 if $i = I$ and 0 otherwise, and λ_{dim} is the learning rate for dimensional attractiveness. Attention is thus transferred between dimensions following each trial on which feedback reveals the model's output to be erroneous; the dimensional attention values, α_i, were defined in Equation 1. For simplicity, we have assumed here that attention learning is a one-step process; a more complex but possibly more accurate approximation may be obtainable by first shifting dimensional attention (much as expert strengths were shifted in Equation 10) and then learning dimensional attractiveness (e.g., Kruschke, 2001b; Kruschke & Johansen, 1999).

Summary

On each trial, a stimulus is presented and the experts each voice an opinion about their preferred response. In parallel, instances in memory are activated on the basis of the extent of their (dimensional attention-sensitive) similarity to the stimulus. The response of one expert is chosen on the basis of stimulus-independent biases and stimulus-specific weights from the activated instances. No blending or averaging of candidate responses takes place. When feedback is made available, the extent of error between each expert's candidate response and the target value is used to adjust the relative strengths of all experts and expert-specific biases and weights between instances and experts.

APPLICATION OF THE THEORY TO DATA

Knowledge Partitioning

Because it was developed to account for knowledge partitioning during function learning, we first apply POLE to the data from Lewandowsky et al. (2002). Figures 2A and 2B show the results of their Experiment 2. As discussed above, in the randomized condition (see Figure 2A), participants' mean responses corresponded to the U-shaped training function, whereas in the partitioning condition, people produced the context-specific extrapolations (and hence the cross-over X pattern) shown in Figure 2B.

According to POLE, participants must partition their knowledge in order to learn any nonlinear function. On that basis, participants in the randomized condition appeared to partition the task accord-

[1] The choice of c as a fixed rather than learned value is arbitrary; it would have been possible to derive a learning rule for c from the overall error gradient.

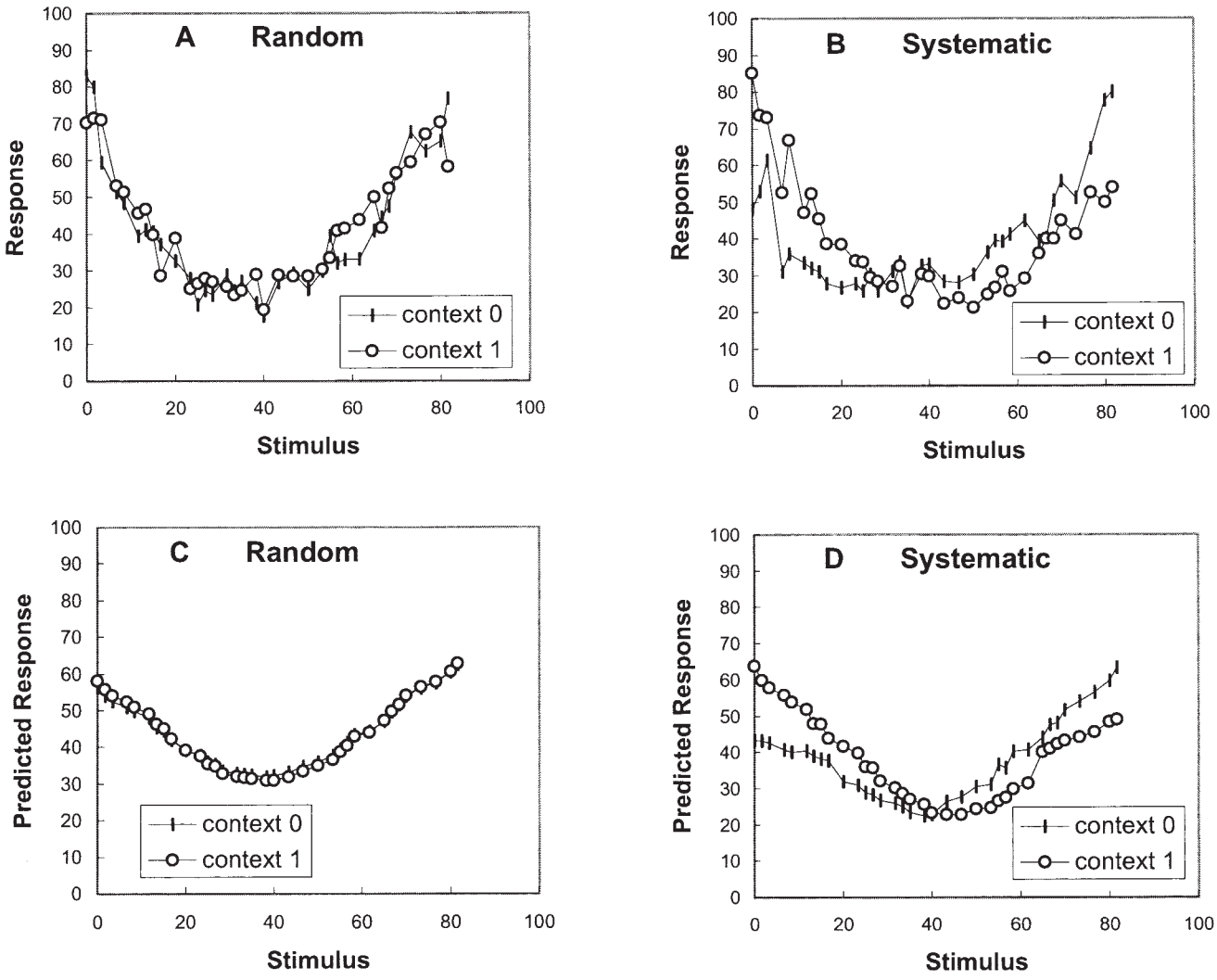

Figure 2. The fit of POLE (population of linear experts model) to Experiment 2 of Lewandowsky et al. (2002). Observed responses (A and B) and POLE's predictions (C and D). A and C are for the random context; B and D are for the systematic context.

ing to stimulus magnitude, whereas those in the partitioning condition used the context dimension. To fit these data, POLE was presented with the same random training sequences seen by participants in the experiment. Both stimuli and responses were rescaled to a unit interval. There were 19 test stimuli presented in each of the two contexts, and the mean response across participants was recorded for each stimulus. The model's mean response was computed analogously, and goodness of fit was determined by the mean-squared deviations between the predicted and the observed means. The model had five free parameters, held constant across conditions: specificity of instance nodes (c); the shift rate for the expert strengths (η_s); and the three learning rates for bias, associative weights, and dimensional attention (λ_b, λ_w, and λ_{dim}, respectively). Because of the small number of test items, POLE was fit to the group data rather than to individual responses.

Figures 2C and 2D show that the model reproduced the knowledge partitioning in the data, accounting for 87% of the variance in the observed mean responses. The best-fitting parameter estimates for this simulation, and all others reported in this article, are presented in Table 1. More important than the variance accounted for, the model captured the critical qualitative form of the data. Participants in the randomized condition learned a single quadratic concept, which POLE represents as a piecewise linear function, whereas participants in the partitioning condition learned two distinct function concepts. POLE correctly segregates the stimuli according to context in this condition, producing two distinct quadratic response functions.

The model accounts for the data by relying on three principles. First, the instance-based representation allows attention to be dynamically reallocated between dimensions during learning. This is necessary for the model to be able to differentiate between stimuli presented in the two training contexts. Second, an entire function, rather than a particular response magnitude, is associated with each stimulus during training. This allows the model to capture the fact that context does not merely shift a learned response but instead alters the relationship between stimulus and response. Third, the

Table 1
The Parameters of POLE for All Experiments

Experiment	c	η_s	λ_w	λ_b	λ_{dim}	ω	ε	Fit R^2	Fit B
Knowledge partitioning (Lewandowsky et al., 2000)	7.85	6.37	0.26	0.40	1.76	0.77	0.29		
Random								.900	
Systematic								.860	
DeLosh et al. (1997)									
Linear	100.00	0.07	0.00	1.01	—	0.40	3.97	.988	
Quadratic	7.27	1.09	1.56	1.00	—	1.55	9.88	.968	
Exponential	90.00	3.06	3.23	0.11	—	1.24	2.25	.984	
Experiment 1	40.97	29.02	0.10	0.36	—	0.01	1.02		38.87[a]
Experiment 2	19.50	30.50	0.11	0.11	—	0.13	18.40		31.80[b]
Experiment 3	32.47	2.42	0.60	0.10	—	0.01	8.34		34.15[c]

Note. Dashes indicate the attention learning parameter was not applicable because stimuli were one-dimensional. POLE = population of linear experts model.
[a] $SD = 5.11$. [b] $SD = 5.33$. [c] $SD = 7.38$.

model selects responses probabilistically, rather than by blending candidates. This final property is critical to the predictions tested by the three experiments below.

We next address the question of whether the architecture of POLE, stimulated by knowledge partitioning, can accommodate benchmark results in function learning. If so, then it may be that function learning in general involves not the integration of knowledge but instead its progressive partitioning.

Assessing the Scope of the Model: Benchmark Results

Busemeyer et al. (1997) identified 10 fundamental empirical principles of function learning that any model must address:

1. Arbitrary associations are harder to learn (i.e., take longer and leave more residual error after fixed training) than are systematic continuous functions.
2. Positive (increasing) functions are easier than negative (decreasing) functions.
3. Strictly monotonic functions are easier than nonmonotonic functions.
4. Cyclic functions, in which more than 1/2 cycle is to be learned, are more difficult than noncyclic functions with only a single inflection point.
5. Linearly increasing functions are easier than nonlinearly increasing functions. Linearity is defined in the psychological similarity space of the stimulus dimensions.
6. People expect functional relationships to be linear, as revealed by the responses made early in training.
7. Interpolation between training stimuli is nearly as accurate as performance on trained stimulus magnitudes.
8. Extrapolation performance is worse than interpolation.
9. The cue labels with which a function is presented for training affect ease of learning: Functions are learned more slowly if labels are incongruent with expectation (e.g., a negative function is difficult to learn if x and y are labeled as "number of drinks consumed" and "blood alcohol level," respectively).
10. Learning of difficult functions (i.e., nonmonotonic or cyclic) is facilitated if stimuli are presented in systematic order, for example from smallest to largest magnitudes.

Benchmark Results: POLE's Account

To examine POLE's inherent ability to predict the known ordering of difficulty among the various functions, we trained randomly parameterized POLE models on seven different functions and analyzed the frequency distribution of the predicted rank orderings of difficulty of the functions. The seven functions were (a) a random mapping between stimuli and responses, chosen such that each value of x had a unique y in the range 0–100; (b) a linearly increasing function ($y = x$); (c) a linearly decreasing function ($y = 100 - x$); (d) a monotonic increasing function ($y = .01 \times x^2$); (e) a monotonic nondecreasing cubic function [$y = 50 + (x - 50)^3/2{,}450$]; (f) a nonmonotonic quadratic function [$y = 1 + .04 \times (50 - x)^2$]; and finally, (g) a cyclic function [$y = 50 + 8.03 \times \sin(x + 7)$].

Sixty stimuli with x ranging from 21 to 80 were generated from each function for use as training instances. Each stimulus was presented four times. The final block was a transfer test, which included a subset of the training items with x ranging from 22 to 80 in steps of two, along with new extrapolation items from the ranges 1–20 and 81–100 and new interpolation items ranging from 21.5 to 79.5 in steps of two. Each simulated experiment, defined as one set of parameter values, involved 10 "participants" trained on the same stimuli but with different random sequences.

We simulated 20,000 experiments by selecting random values for the six model parameters from the following ranges: $c \in (1, 100)$, $\eta_s \in (0, 20)$, $\lambda_w \in (0, 2)$, $\lambda_b \in (0, 2)$, $\varepsilon \in (0, 10)$, and $\omega \in (0, 10)$. Because the to-be-learned functions were all one-dimensional, there is no dimensional attention learning parameter. For each experiment, the mean absolute error at transfer was

computed for each of the three classes of test stimuli (old training instances, novel interpolation instances, and novel extrapolation instances). Performance on the old instances represented difficulty of learning and was used to rank order the seven functions. There are 5,040 different possible orderings of the seven functions, of which 360 were observed in the simulations. Analysis of Principles 1–6 focused on the frequency distribution of those 360 orderings among the 20,000 simulated experiments.

Principles 1 and 4

Cyclic functions are harder than any save random; on only 2.1% of parameter settings was any function other than random harder to learn than the cyclic. On 76.5% of cases, the cyclic was learned better than the random function. The difficulty with cyclic and random functions is due to the graded similarity of the instance-based representations POLE uses to choose among the experts. It is much easier for the model to learn functions where y changes smoothly as x varies (and adjacent stimuli thus tend to reinforce each other), than the random function and, to a lesser extent, the cyclic function (where adjacent stimulus magnitudes may require competing responses).

Principles 2 and 6

Positive linear functions are easier to learn than negative linear functions. This is true for the model only if the preferred slope is positive and if the initial bias and the bias decay (ω and ϵ in Equation 4) are both large. The preferred slope was always set to 1.0, and when all other parameters were freely manipulated, just over 92% of replications showed a preference for positive over negative functions.

Principle 3

A total of 98.5% of parameters resulted in nonlinear, monotonically increasing functions being easier than nonmonotonic (but noncyclic) functions. This difference is due both to the rate-of-change sensitivity that captures Principles 1 and 4 and to the strength of initial biases in favor of positive linear functions, which are more similar to the monotonically increasing functions than to the nonmonotonic functions.

Principle 5

People find linear functions easier to learn than nonlinear functions. This was captured by POLE, as revealed by the following pairwise partial orderings: The positive linear function was easier to learn than was the (nonlinear) monotonically increasing function in over 98.9% of all parameter settings; it was easier than the (nonlinear) monotonically nondecreasing function 92.7% of the time, easier than the nonmonotonic in 99.3% of all cases, easier than the cyclic in 99.98%, and easier than the random in 99.97% of all parameter settings.

The most common 5 of the 360 observed orderings accounted for 71.7% of all results, and all of these 5 subsumed the ordinal relationships observed in the data: namely (positive linear) $<$ (negative linear) $<$ (monotonic increasing $\cap$ monotonic nondecreasing $\cap$ nonmonotonic) $<$ (cyclic $\cap$ random).

Principles 7 and 8

Analysis of the simulated experiments confirmed Principle 7 (interpolation accuracy equals that for trained items) and Principle 8 (extrapolation is not as accurate as interpolation). Extrapolation was better than interpolation on only 4.25% of all cases, with most (85%) of these being for the negative linear function. Similarly, interpolation accuracy could not be reliably distinguished from accuracy on old instances; on 56% of cases accuracy was better on old items, and 44% of the time it was better on interpolation items. The mean advantage in accuracy for old items was only 0.05 points (responses were measured on a 100 point scale), with a variance of 0.09 points.

Summary

Our simulation analysis of the model's parameter space suggests that POLE generally makes the correct qualitative predictions about function learning across a range of experimental conditions (different sequences, functions, and parameters). It follows that POLE's basic model architecture, and not its parameters, permits it to handle many of the existing empirical principles.

It appears that these empirical principles reflect two underlying regularities. First is that difficulty of learning is strongly correlated with a function's complexity. POLE learns complex (nonlinear) functions by piecewise linear approximation. The more pieces are required, the longer learning will take. The second regularity is that people are predisposed to learn some functions (positive linear) faster or better than others (negative linear) even when complexity is equivalent. This is embodied in the way POLE begins each experiment with a distribution of bias strengths that favor positive linear functions. In the General Discussion, we take up Principles 9 and 10 in relation to this aspect of POLE.

Benchmarks: Specific Experiments

Although POLE qualitatively captures the benchmark results of function learning, it does not necessarily follow that POLE is equally suited for quantitative modeling of detailed existing results. DeLosh et al. (1997) presented several experiments that emphasized Principle 8 (the difference between interpolation and extrapolation) and that were accompanied by a successful application of the EXAM model. Could that account be rivaled by POLE?

DeLosh et al. (1997) trained groups of participants on linearly increasing, nonlinearly increasing (exponential), or nonmonotonic (quadratic convex) functions. The average responses to extrapolation items tended to resemble linear extrapolation from the nearest two or three training items, as shown in Figures 3A–3C. Figures 3D–3F also show that EXAM, with its assumption of linear extrapolation from closest trained neighbors, fit these data quite well. We obtained the EXAM predictions reported in this figure by fitting the model to the transfer data because we did not have the trial-by-trial learning data used by DeLosh et al. EXAM was trained on a set of 25 randomly generated training sequences, and the model's performance index was the average of these 25 replications.

We fit POLE to the data in the same way, with 25 participants per function, optimizing parameters for each function separately by reducing the difference between obtained and predicted (transfer) responses. POLE accounted for all aspects of the data, including the generally linear extrapolation performance and the more accurate learning of the linear function as compared with the other

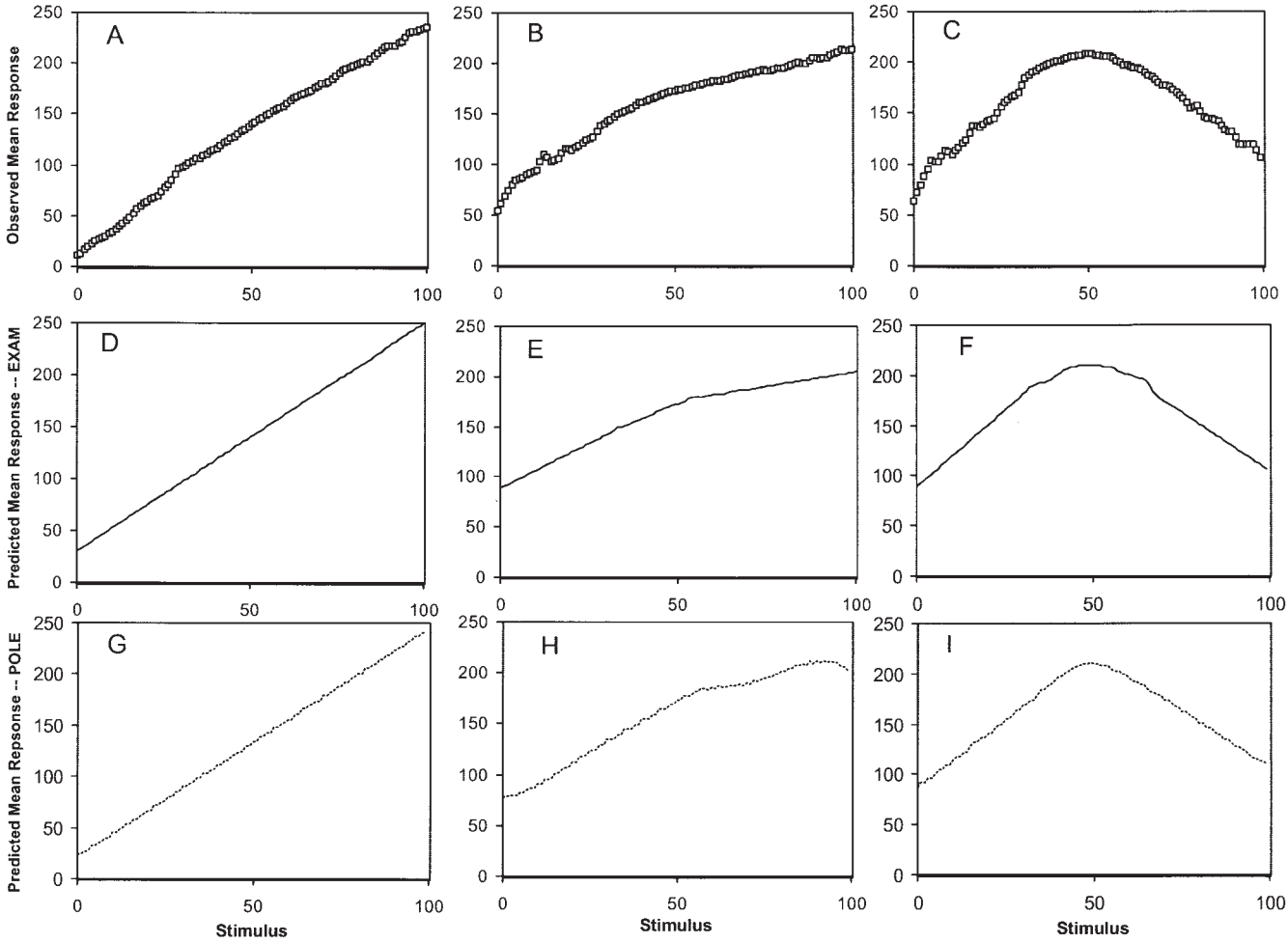

Figure 3. The results of DeLosh et al.'s (1997) experiments, fit by EXAM (extrapolation–association model) and POLE (population of linear experts model). A–C: Observed responses after training with linear, exponential, and quadratic functions, respectively. D–F: EXAM's predicted responses. G–I: POLE's predicted responses.

two function concepts (see Figures 3G–3I for predictions and Table 1 for parameter values).

POLE: A Quantitative Account of the Existing Literature

POLE accounts for a wide range of function learning phenomena. POLE handles benchmark results because of its basic architecture rather than specific parameter values, and it can quantitatively capture crucial experimental results (e.g., DeLosh et al., 1997). Moreover, unlike EXAM, POLE handles the knowledge partitioning observed by Lewandowsky et al. (2002). POLE also makes novel predictions about function concepts; having underscored the model's utility, we now explore some of these predictions.

PARTITIONING WITHOUT CONTEXT: ENGENDERING MULTIMODAL RESPONSES

The principal assumption of POLE, that people learn to associate purely linear response functions with partial stimulus ranges, gives rise to a counterintuitive prediction. Because partitioned learning always occurs in POLE—unless a single linear function is adopted—any uncertainty about which expert should be used to generate a response will express itself as multimodality of the response distribution.

For example, when a quadratic function is learned, uncertainty about the appropriate expert might arise at the boundaries between linear segments. In most cases, including the ones considered thus far, this uncertainty escapes experimental detection because the alternative functions provide very similar response magnitudes in the vicinity of their splice. However, if uncertainty about the appropriate expert could be experimentally induced in regions of the stimulus space in which candidates yield widely divergent responses, then response magnitudes should exhibit systematic multimodality. That is, for an identical stimulus, response magnitudes should cluster around the values provided by the various competing experts. This prediction is unique to POLE and, as we confirm by simulation later, cannot be handled by EXAM. We now report three experiments that sought to induce multimodality by introducing various discontinuities into the to-be-learned function.

Experiment 1

The to-be-learned function in the first experiment consisted of two linear segments with different intercepts but equal slope, spanning different ranges (see Figure 4). For the lower component, $x \in (20, 45)$; for the upper component, $x \in (55, 80)$.

The two linear segments were separated by a set of stimulus magnitudes that were not shown during training and were presented for the first time during a transfer test. One possible transfer pattern would be for people to interpolate linearly between the closest training stimuli, thus giving rise to a quasi-sigmoid and relatively smooth overall function. The response pattern predicted by POLE, conversely, is that the absence of training in the gap between linear segments causes uncertainty about which response function to choose. In consequence, for each such stimulus, people would be expected to choose one or the other learned linear response segment, either at random or on some other preferential basis. Across repeated presentations of a given stimulus, the independent and parallel extrapolations of each linear segment should, in turn, produce distinct bimodality.

Method

Participants

Participants were 39 members of the campus community at the University of Western Australia ($n = 13$) and Indiana University ($n = 26$) who participated voluntarily. Participants either received course credit or were remunerated at the rate of $5/hr.

Apparatus and Stimuli

The experiment was controlled by a PC-compatible computer that presented stimuli and collected responses. The to-be-learned function shown in Figure 4 was arbitrarily parameterized as $y = .7\,x$ (for $0 < x < 50$) and $y = 30 + .7\,x$ (for $50 < x < 100$). In this study, as in all remaining experiments, both variables (x and y) were scaled to range from 0 to 100 logical units. All stimuli and analyses refer to logical units. Owing to the standard aspect ratio of the monitor, the physical lengths of the response scale (which was vertically oriented) differed from that of the (horizontal) stimulus scale. On a 15-in. monitor, 100 units of x subtended 15.66 cm, whereas 100 units of y corresponded to 12.25 cm.

Training stimuli were all integer values of x in the ranges 20–45 and 55–80, resulting in a total of 52 unique training items. Transfer stimuli were all integer values of x within the ranges 6–19, 46–54, and 80–95, resulting in a set of 39 transfer items.

Each stimulus involved graphical elements only, without any numeric or textual labeling. The stimulus variable (x) was represented by the length of a red horizontal bar at the top of the screen. The 12.25 cm high response scale was located below and to the right of the stimulus bar and appeared simultaneously with it. To select a response magnitude, participants adjusted a scrollbar using the mouse from a point of origin at the midpoint of the scale. Once a value was selected, participants clicked on a button to register their response. No time limit was imposed on selecting a response.

On training trials, if the response was within ± 4 units of the correct value for y, no explicit feedback was presented, and the next trial commenced after a delay of 2 s. If a response deviated by more than 4 units (approximately 0.5 cm on a 15-in. monitor), the correct value of y was indicated by a vertical bar that appeared next to the response scale. Participants then had to adjust the scrollbar in response to the feedback until they indicated the correct value. The next trial commenced after a delay of 2 s. On transfer trials, a response was followed by the message "no feedback available," which remained on the screen for 2 s until the next trial commenced.

Procedure

The experiment consisted of four blocks of trials. The first three blocks involved only training trials, and the final block involved both training and transfer trials. All 52 training items were presented once within each block, and each of the 39 transfer items was presented twice during the final block, yielding a total of 286 experimental trials. The order of trials was random within each block, and a different random sequence was used for each participant.

Instructions to participants emphasized that the relationship between x and y was arbitrary and that the experimental task required learning of that relationship. Participants were informed before the final block that some trials would not include feedback. Experimental sessions lasted about 40 min.

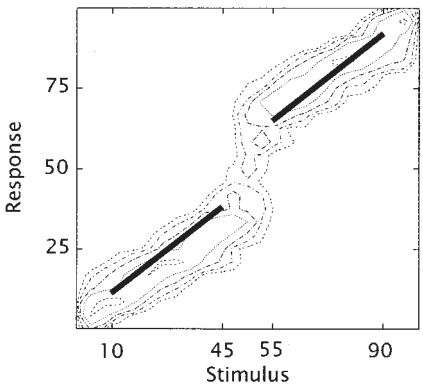

Figure 4. Observed frequency distribution (contour plot) of responses for the last block (including both training and transfer items) and training stimuli (thick lines) for Experiment 1. Both parts of the to-be-learned function were presented with equal frequency during training. Results are aggregated across all 30 participants and transformed into relative frequencies. Note the bimodality of responses in the central ($45 < x < 55$) transfer region. In this and all subsequent plots of conditional frequency distributions, isocontours are marked at .05, .10, .20, .40, and .80.

Results and Discussion

Training Performance

Training performance was measured by the average absolute deviations between the true function values and response magnitudes to the 52 training stimuli across the four blocks. For each participant, deviations were averaged across training stimuli within each linear segment separately. The means of those deviations are shown in Table 2 as a function of blocks of trials.

It is clear from the data that people very rapidly learned to produce the correct response magnitude for both function segments with little error. This was confirmed by a 2 (function segment) × 4 (training block) within-participants analysis of variance (ANOVA), which yielded a highly significant main effect of training block, $F(3, 114) = 29.39$, $MSE = 3.24$, $p < .01$, no main effect of function segment, $F(1, 38) = 2.63$, $p > .10$, and a marginal interaction between the two factors, $F(3, 114) = 2.53$, $MSE = 1.54$, $p = .06$. The marginal interaction reflected the slight, selective decrease in performance during the final block for the lower function segment. We do not pursue this result further.

Table 2
Performance During Training, Measured as Mean Absolute Error, for Participants in Experiments 1–3

Experiment and condition	Block 1	Block 2	Block 3	Block 4
1				
Upper half	6.61	4.23	3.89	5.15
Lower half	6.18	4.14	4.00	4.24
2				
Positive	8.55	7.13	6.15	6.81
Negative	31.81	29.31	25.88	27.07
3				
Positive	8.84	7.47	7.31	7.15
Exceptions	30.58	25.87	24.04	25.81

Transfer Performance

Of greatest interest here was the predicted occurrence of bimodality in the transfer region between the two linear segments. The transfer data are shown in Figure 4 as a frequency contour plot of all responses made by participants in the last block of the experimental session. The critical transfer items were only presented in the last training block. The pattern in the figure remains largely unchanged if the first three training blocks are included.

The figure suggests considerable heterogeneity among responses in the critical central transfer region with little evidence for a smooth or consistent interpolation between the two linear segments. The remaining analyses focused on statistical confirmation of the multimodality of the transfer data. The data were analyzed first at the aggregate level and then at the level of individual participants.

Aggregate analysis. The aggregate analysis was performed by considering all participants' responses in the central transfer region. There were 699 such responses (3 observations were missing). To permit aggregation across magnitudes of x, we rescaled each response as a residual from the y value obtained from the lower linear segment.[2] The cumulative frequency distribution of those residuals, with 64 bins of unit width, was subjected to a dip test for bimodality (Hartigan & Hartigan, 1985). The value of the dip statistic (0.078) was found to be highly significant at $p < .01$. The significant effect confirms the pattern suggested in Figure 4, in that responses do not occur in the region between the two dominant responses, and establishes the presence of bimodality in the transfer data, as predicted by POLE.

Analysis of individual participants. One limitation of the preceding analysis is that it cannot determine whether the observed bimodality reflects different participants or different responses by the same individuals. The bimodality may have arisen because some participants exclusively used the lower function segment to extrapolate into the transfer region and others exclusively used the higher function. The unique prediction of POLE, alternatively, is that the same individual may, on different occasions, choose one or the other expert to make a response to the same test item.

The presence of bimodality within participants was confirmed by a comparison of the consistency of a person's transfer responses between the critical central region (i.e., $46 < x < 54$) and the extrapolation regions at the lower ($x < 20$) and upper ($x \geq 80$) extremes. For each participant, differences were computed between response magnitudes across the two repeated presentations of each transfer stimulus. These differences were averaged separately across the 9 transfer items within the critical central region and across the 30 transfer items in the extreme extrapolation regions. The mean difference across participants was 12.34 units for the critical central region and 4.64 units for the extreme extrapolation regions. The effect of transfer region was found to be highly significant, $t(38) = 8.14$, $p < .01$.

Because this analysis contrasted participant's transfer responses between two situations, only one of which involved uncertainty about the choice of expert, the outcome confirms that bimodality is not an inherent characteristic of any set of transfer responses but is limited to those situations in which it is predicted by POLE. Figure 5 shows 4 representative participants' responses during the experiment. We now examine the ability of EXAM and POLE to account for the responses of those individuals.

Theoretical Analysis

Fit of EXAM

EXAM was implemented as described in Appendix A. The model was applied to the data from Experiment 1 in three stages. First, EXAM was fit separately to the trial-by-trial responses of 2 representative participants by minimizing the root-mean-squared deviations (RMSDs) between the model's mean predicted response (see Equation A10) and each participant's responses. Fitting of 2 participants, 1 of whom was strikingly multimodal, was considered sufficient because if EXAM cannot produce multimodal responses for that participant, it would be pointless to consider others. All weights were initialized to 0 at the outset.

The results are shown in Figure 6. Figure 6A shows data and predictions for 1 of the few participants (Participant 1 from Figure 5) who was visually identified as interpolating smoothly between function segments. Figure 6B shows the corresponding results for a participant (Participant 4 from Figure 5) who exhibited bimodality in the central transfer region. It is clear from Figure 6 that EXAM provided a very satisfactory account (RMSD = 5.16) for the smoothly interpolating participant with two parameters. (The best-fitting parameter values were $\gamma = 0.54$, $\eta = 0.05$; see Appendix A for an explanation of parameters.) However, it is also clear that EXAM cannot handle the bimodality exhibited by the other participant, as reflected by the larger RMSD of 9.89 ($\gamma = 0.99$, $\eta = 0.01$).

Exploration of the parameter space confirmed that EXAM could not accommodate the bimodal participant across a wide range (0.01–2.00) of settings of the generalization parameter γ. The only way EXAM can produce different responses to the same stimulus is if training trials intervene and learning were to change the association weights within the model. Thus, EXAM's failure to produce systematic bimodality is perhaps not surprising.

To give EXAM every possible opportunity to generate multimodality, we modified it (see Appendix A) to respond to the same stimulus with any of a number of different response magnitudes, each with a predicted probability of occurrence. We then fit the data of the representative bimodal participant using the probability distribution of possible responses introduced in Appendix A as the

[2] The choice of the lower linear segment for computation of residuals is arbitrary. The results are unchanged if the upper segment is used instead as all this does is recenter the distribution on a different mean.

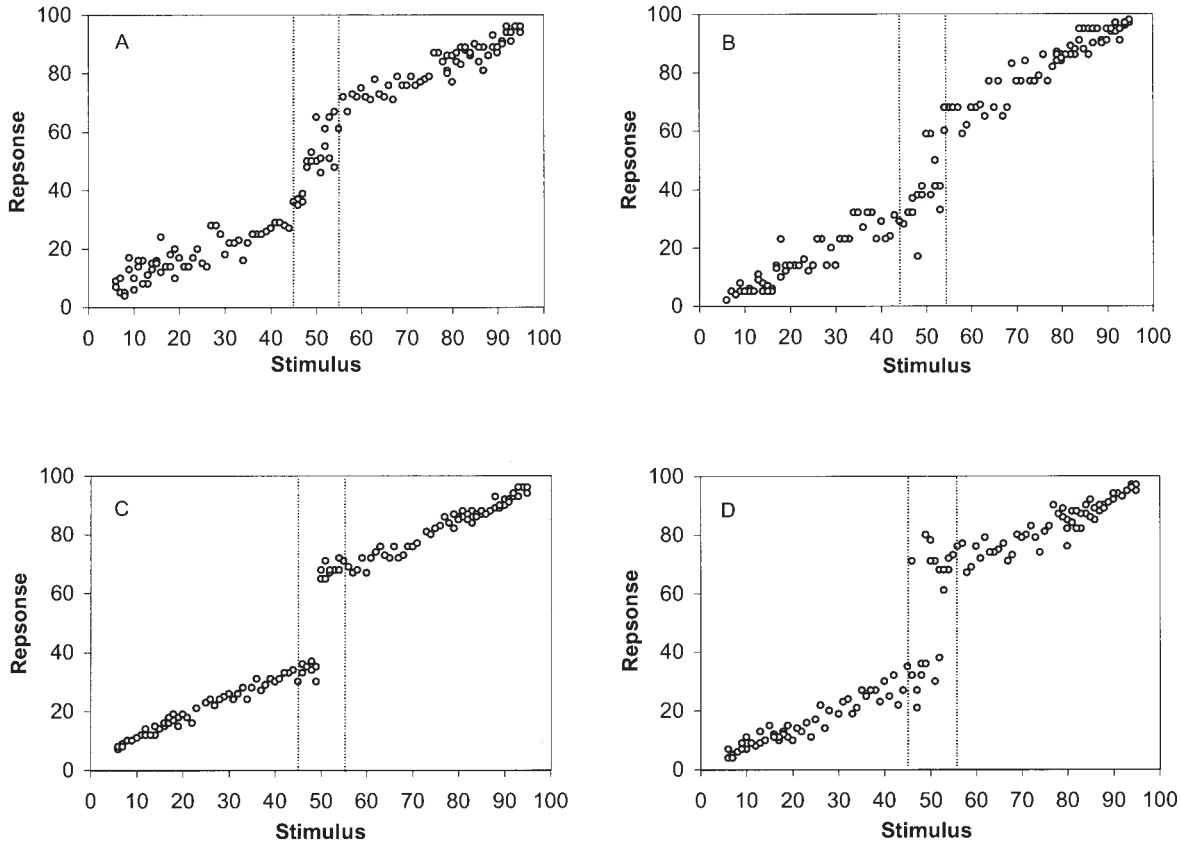

Figure 5. Responses of 4 participants in Experiment 1, showing the diversity of response patterns to the final block of stimuli. The dashed vertical lines mark the bounds of the training stimuli; stimuli between the lines were presented without feedback only. Dip scores were 0.17, 0.21, 0.13, and 0.11 for Participants 1–4 (shown in A–D), respectively.

model's predictions (see Equation A10). The badness-of-fit index, *B*, used to fit response distributions is defined in Appendix C; see in particular Equation C2. Figure 7A shows a probability-contour plot of the best-fitting predicted distribution of response magnitudes ($B = 432.03$, $\gamma = 1.98$, $\eta = 0.05$). It is clear that EXAM does not reproduce the bimodality exhibited by the participant (whose data are replotted as a frequency-contour plot in Figure 7D), even when provided with the opportunity to nominate multiple response candidates.

Further exploration of the parameter space revealed that EXAM could predict a limited extent of bimodality only when the slope of the Gaussian generalization gradient was extremely shallow ($\gamma =$

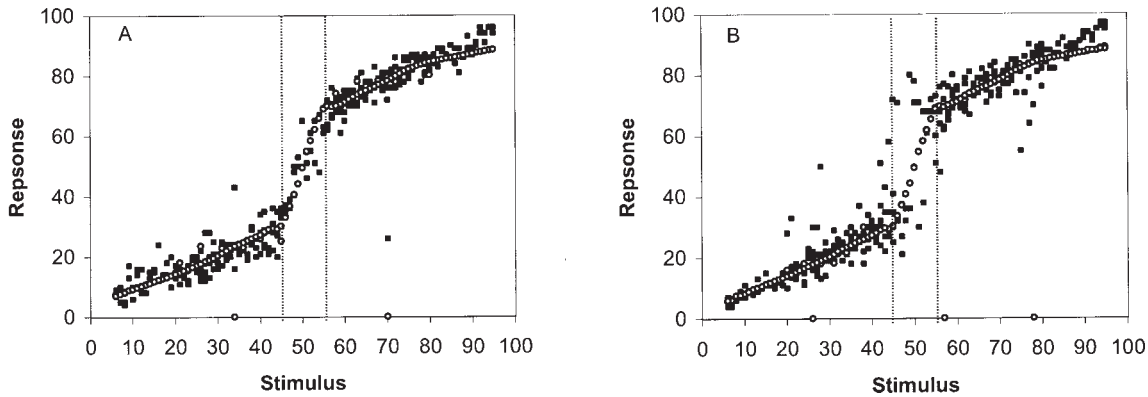

Figure 6. The best fit of EXAM (extrapolation–association model) to 2 participants from Figure 5. Dashed vertical lines mark the bounds of the training stimuli. A: Participant 2. B: Participant 4. The fit to each participant is based on root-mean-squared deviation between mean response and data.

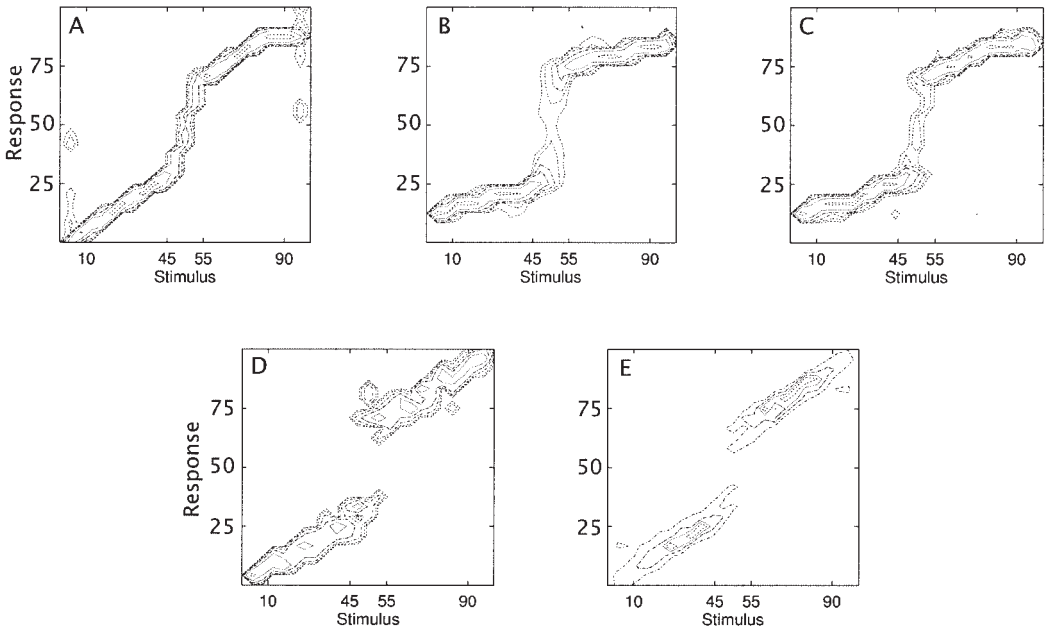

Figure 7. EXAM (extrapolation–association model) cannot produce appropriate bimodality, whereas POLE (population of linear experts model) can (data from Participant 4 of Figure 5). A: A contour plot of EXAM's predicted response probabilities produced from the best fit using log-likelihoods. B: The maximum obtainable bimodality. C: The best-fitting predicted response probabilities with preloading of weights. D: The data replotted as a frequency-contour plot for comparison. E: POLE's best-fitting predicted probabilities.

0.01). This limited bimodality was stable across two settings of the learning parameter, $\eta = 0.05$ and $\eta = 0.10$. Figure 7B shows the predicted probability contour for $\eta = 0.05$.

We argued earlier that people's prior expectations of a positive linear function might be critical to certain aspects of knowledge partitioning, by providing competition between preexisting and learned response tendencies. Appendix A describes how EXAM was provided with preinitialized weights to reflect this prior expectation. Figure 7C shows the best-fitting probability contour for this three-parameter model ($B = 187.31$, $\gamma = 0.03$, $\eta = 0.27$, $\beta = 1.81$). Although the modeling of prior expectations reduced the badness of fit considerably, the figure shows that EXAM was unable to predict the observed bimodality despite the inclusion of a third free parameter.

Fit of POLE

POLE was fit to the responses from each individual in the same manner as EXAM. In sharp contrast to EXAM, the model captured the diversity of response types quite well, producing predictions of both multimodal and unimodal responses for different participants. Figure 7E shows the response probabilities predicted for Participant 4 by the best-fitting parameters of POLE ($c = 13.99$, $\eta_s = 4.19$, $\lambda_w = 0.31$, $\lambda_b = 0.001$, $\omega = 0.21$, $\epsilon = 0.00$, and $B = 23.57$). The panel clearly demonstrates that multimodality in the central region is quite within POLE's predictive ability.

The individual-participants analysis is necessary because of our concern with intraparticipant variability. However, the fits also identified significant commonalities between participants as revealed by a principal-components analysis of the 39 sets of estimated parameter values. The principal-components solution discovered only two components with eigenvalues greater than one, which together accounted for over 70% of the variance. A parallel cluster analysis revealed only a single dominant cluster.

Thus, despite clear individual differences, it appears as though the mean parameter values might represent a good approximation of the parameters estimated for each participant. When fit to each participant separately, the mean B was 38.00 ($SD = 5.30$). Given the coherence of the distribution of parameter values, we chose to follow the individual-participants analysis with a constrained optimization of the model, which provided POLE with only six free parameters total, rather than six per participant. These parameters were optimized against the trial-by-trial responses of each individual participant, as described in Appendix C. The best-fitting six parameters (listed in Table 1) produced a mean B of 38.87 ($SD = 5.11$). This implies that constraining POLE to have six parameters total instead of six per participant (for a total of 228) reduced fit by only 2.3%, or only 17% of a standard deviation. Figure 8 shows the predicted frequency distribution derived from the single parameter set. Again, bimodality in the central region is clearly present, whereas responses are otherwise largely governed by the two dominant response functions $y = .7x$ and $y = 30 + .7x$.

Experiment 2

The first study inserted a gap between two vertically offset linear function segments to induce uncertainty about the choice of response function. The second experiment extended this approach by presenting people with a positive linear function that contained three gaps. Each gap, in turn, included one training stimulus in its center that was presented repeatedly and that required an excep-

tional response not accommodated by the linear function. The resultant to-be-learned function is shown in Figure 9.

The to-be-learned function has three critical features: First, because the *x* values of the exception stimuli differed from their closest neighbors on the linear function by several units, the overall function was in principle learnable given sufficient discrimination among stimulus magnitudes. Second, the three exception stimuli jointly defined their own negative linear function. Third, and perhaps most important, the training stimulus with the largest *x* value required an exceptional response, thus posing a particular challenge during extrapolation. In particular, the local slope estimates that underlie extrapolation in EXAM are always dominated by the training item most similar to any given test item. Visually, the to-be-learned items suggest a positive linear function. EXAM, however, learns only associations of response magnitudes, not functions, and so the final exception item must cause EXAM to always extrapolate off the dominant training function. POLE, by contrast, predicts that under these circumstances people learn the two opposing functions simultaneously, either by associating each with a different set of stimulus magnitudes or by adopting one function as a default and learning the other by association with exception items.

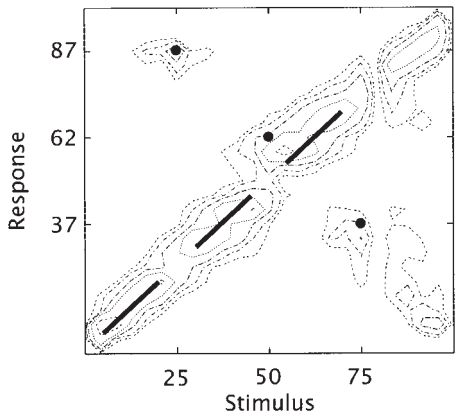

Figure 9. Observed frequency distribution (contour plot) of responses for the last block (including both training and transfer items) and training stimuli (thick lines) for Experiment 2. Training items on the negative diagonal (shown as solid circles) were presented with higher frequency than items on the positive diagonal. Results are aggregated across all 32 participants. Note the strong bimodality in the transfer region ($x > 80$).

Method

Participants

Participants were 32 members of the campus community at the University of Western Australia ($n = 7$) and Indiana University ($n = 25$) who participated voluntarily. Participants either received course credit or were remunerated at the rate of $5/hr.

Apparatus and Stimuli

The experiment was controlled by a PC-compatible computer that presented stimuli and collected responses. The positive to-be-learned function shown in Figure 9 was arbitrarily parameterized as $y = x$.

Training stimuli consisted of two sets. First, all integer values of *x* within the ranges 6–20, 30–45, and 55–70 were sampled from the function $y = x$ to form the set of 48 *positive* training items. In addition, there were three *exceptional* stimuli with *x* values of 25, 50, and 75, with associated response magnitudes of 87, 62, and 37, respectively. Exception stimuli thus fell on the negative linear function $y = 112 - x$.

Transfer stimuli were all integer values of *x* within the range 81–94, thus resulting in 14 transfer stimuli.[3] Stimulus presentation and response details were the same as in Experiment 1.

Procedure

The experiment consisted of four blocks of trials. The first three blocks involved only training trials, and the final block involved both training and transfer trials. The 48 positive training items were presented once within each block, and the three exceptional stimuli were presented 8 times each, yielding 72 training trials per block. The final block additionally included two presentations of each of the 14 transfer items, yielding a total of 316 experimental trials. Order of trials within each block was randomized anew for each participant. All remaining procedural details were the same as in Experiment 1.

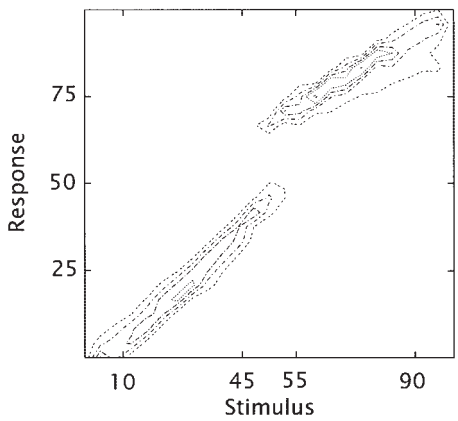

Figure 8. The predicted distribution of response frequencies for the last block of trials in Experiment 1, as the result of the single best-fitting set of parameters for POLE (population of linear experts model; see Table 1). The bimodality in the central region is similar to that seen in the data (see Figure 4).

Results and Discussion

Training Performance

The analysis considered the 48 positive training stimuli and the 3 exceptional stimuli separately. Table 2 shows average performance for both classes of stimuli across blocks.

The data reveal several clear findings. First, there is evidence of learning for both classes of stimuli. Second, and not unexpectedly, performance on the exceptional stimuli remained vastly inferior to performance on the predominant positive training items. Both of these effects were confirmed by the 2 (type of stimuli) × 4 (training block) within-participants ANOVA, which revealed a main effect of type of stimuli, $F(1, 31) = 163.14$, $MSE = 179.00$,

[3] The University of Western Australia participants actually received 15 transfer trials (including $x = 80$). However, because the $x = 80$ stimulus was lost because of a file transmission error and Indiana participants thus only received 14 transfer trials, we restrict analysis to those 14 trials experienced by all participants.

$p < .01$, and training block, $F(3, 93) = 9.05$, $MSE = 22.82$, $p < .01$, but no interaction between the two variables, $F(3, 93) = 1.55$, $p > .10$. The reduced extent of learning on the exceptional items is consonant with related research in categorization, which also shows that a small number of exceptions to a rule are classified less well than rule-conforming items, even when memorization of the few exceptions could have led to perfect performance (e.g., Lewandowsky et al., 2000).

To examine whether participants learned anything about the exception items, we calculated the mean response to these items in each block of training. We also calculated the response that would be expected for those items in each block on the basis of what people had learned about the positive stimuli alone. Thus, the responses to the positive items were regressed onto stimulus magnitudes in each block, and the regression equation was used to predict what performance would have been on the exception instances had people applied the same response function that they were using for the positive items. Table 3 shows the results of this analysis: It is clear that responses to the exception items increasingly diverge from the responses to other items and move toward the correct responses. This indicates that although participants were not as accurate with exception items as they were with rule-consistent items overall, they were clearly not ignoring the exceptions during learning.

Transfer Performance

The transfer data are shown in Figure 9, which presents a frequency-contour plot of all responses to the final block of the experiment. Statistical support for multimodality was again produced at the aggregate level as well for individual participants.

Aggregate analysis. The available 885 transfer responses from all participants (11 observations were missing) were rescaled as residuals from the positive ($y = x$) function. The cumulative frequency distribution of those residuals, with 97 bins of unit width, was subjected to the dip test. The value of the dip statistic (0.208) was found to be highly significant at $p < .01$, which confirms the bimodality that is evident in Figure 9.

Individual-participants analysis. To confirm the presence of bimodality within participants, the dip test was applied separately to each individual's 28 transfer responses. Of the 32 tests, 15 were significant at the .05 level with a dip statistic in excess of the cutoff of 0.089 ($n = 28$). Thus, nearly half of all participants exhibited statistically detectable bimodality in their transfer responses. Representative individual responses are shown in Figure 10: Figures 10A and 10B show the responses of participants who exhibited the least bimodality, whereas the responses in Figures 10C and 10D are from the participants who exhibited the greatest extent of bimodality.

There is clear evidence, therefore, that a very large proportion of participants responded bimodally at transfer. Moreover, as suggested by Figures 10C and 10D, those participants who exhibited bimodality clearly alternated between use of the positive ($y = x$) and exceptional ($y = 112 - x$) function.

Theoretical Analysis

Fit of EXAM

EXAM was again fit to the responses of a representative participant (Participant 4 from Figure 10) who exhibited strong bimodality. The fit minimized the deviations between the predicted probability distribution of responses and the participant's responses. Figure 11A shows EXAM's best-fitting predicted probability contour ($B = 102.30$, $\gamma = 0.005$, $\eta = 0.08$). Figure 11C shows the responses of the participant replotted from the earlier figure as a frequency-contour plot for visual comparison. Figure 11A clearly shows that EXAM's extrapolations were entirely based on slope estimates that involve the last exception ($x = 75$). Accordingly, EXAM failed to predict the majority of observed extrapolations, which fell along the positive function. The contour plot also clarifies that EXAM cannot predict bimodality at any stimulus magnitude, either within or outside the training range.

Figure 11 also shows that EXAM, with this set of best-fitting parameter estimates, could not predict mean responses for the outlying training stimulus at $x = 25$ ($y = 87$). This occurred because the generalization gradients were extremely wide, thus precluding the acquisition of a single outlying response among neighboring stimulus magnitudes. In confirmation, the outlying response at $x = 25$ could be learned by EXAM if the generalization gradients were narrowed. Specifically, with parameter values $\gamma = 0.97$ and $\eta = 0.04$, EXAM was able to predict mean responses for all three outlying training stimuli; however, under those parameter settings, EXAM also predicted unimodally negative response magnitudes for all extrapolations beyond the training range. Given that participants could not register responses less than 0, these predictions were considered implausible.

The failure of EXAM to handle crucial aspects of these data may have arisen because weights were initialized to 0 at the outset, thereby conceivably preventing use of the positive function for extrapolation. This possibility was ruled out by a fit of the three-parameter version, in which weights were preinitialized to capture

Table 3
The Prediction Equation and the Difference Between Responses to Exception Items and Prediction in Experiment 2

Block	Equation	Exception 1	Exception 2	Exception 3
1	$y = 13.65 + .73x$	32.6 (31.9)	52.1 (50.2)	66.4 (68.4)
2	$y = 8.91 + .82x$	35.4 (29.4)	52.5 (49.9)	59.6 (70.1)
3	$y = 7.40 + .84x$	42.9 (28.4)	52.2 (49.4)	58.3 (70.4)
4	$y = 6.95 + .86x$	43.3 (28.5)	52.1 (50.0)	59.7 (71.5)
Training values	$y = x$	87 (25)	62 (50)	37 (75)

Note. Values in parentheses are predicted values based on the regression equation.

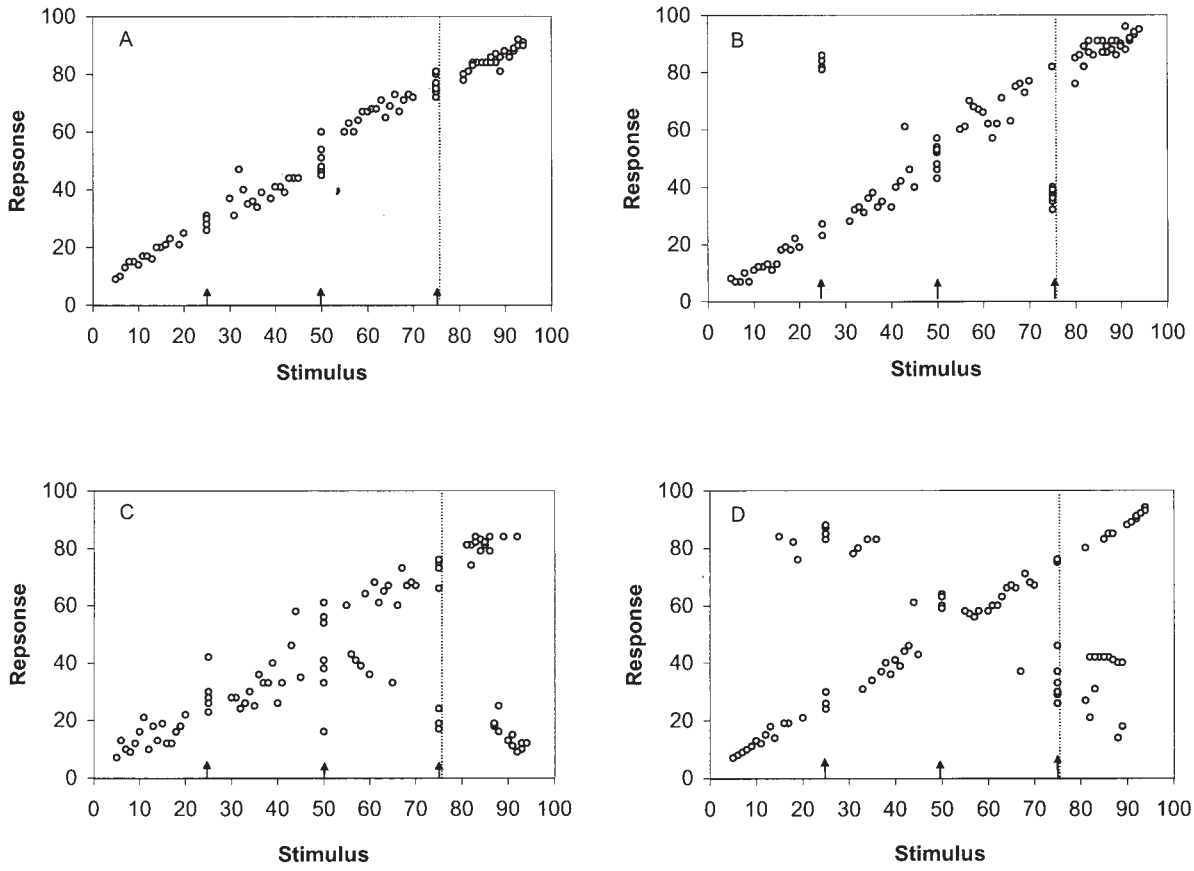

Figure 10. Responses of 4 participants in Experiment 2, showing the diversity of response patterns to the final block of stimuli. The dashed vertical line in each panel marks the boundary of the training stimuli; items to the right of the line were presented without feedback only. The three small arrows on each *x*-axis mark the locations of the high-frequency exception items. Dip scores are 0.047, 0.020, 0.073, and 0.173 for Participants 1–4 (shown in A–D), respectively.

the expected positive linear function. The best-fitting results (B = 97.19, γ = 0.005, η = 0.08, β = .03), shown in Figure 11B, continue to be unimodal, show no learning of the outlying training stimulus, and exhibit no extrapolations along the positive function.

Fit of POLE

POLE was again fit to each participant individually The mean goodness of fit was B = 30.96 (SD = 5.21). POLE's predicted response distribution for 1 participant (with parameters c = 16.72, η_s = 1.00, λ_w = 0.95, λ_b = 0.10, ω = 0.12, ϵ = 10.00, and B = 26.45) is shown in Figure 11D. Unlike EXAM, POLE clearly accommodates the bimodality in the transfer region.

When each participant's data were predicted with only six overall parameters, rather than six for each participant, we obtained a mean fit of 31.80 (SD = 5.33). Thus, the loss of 186 free parameters reduced fit only by 2.6%, or 16% of a standard deviation. The best-fitting six parameters (listed in Table 1) produce a clearly multimodal distribution during the final block, which is shown in Figure 12.

Experiment 3

The second experiment demonstrated clear evidence of bimodality at transfer when people were presented with two interleaved but conflicting functions. The final experiment examined the effects of presenting exceptional training stimuli that did not fall along a linear function. The to-be-learned function for Experiment 3 is shown in Figure 13.

Method

Participants

Participants were 45 members of the campus community at the University of Western Australia (n = 14) and Indiana University (n = 31) who participated voluntarily. Participants either received course credit or were remunerated at the rate of $5/hr.

Stimuli, Apparatus, and Procedure

The only difference to the previous experiment was that the y values for two exceptional stimuli were swapped. The new exceptional stimuli were x = 25, y = 62; x = 50, y = 87; and x = 75, y = 37.

Results and Discussion

Training Performance

Table 2 shows average performance for positive and exceptional stimuli across blocks. The training data resembled those of the

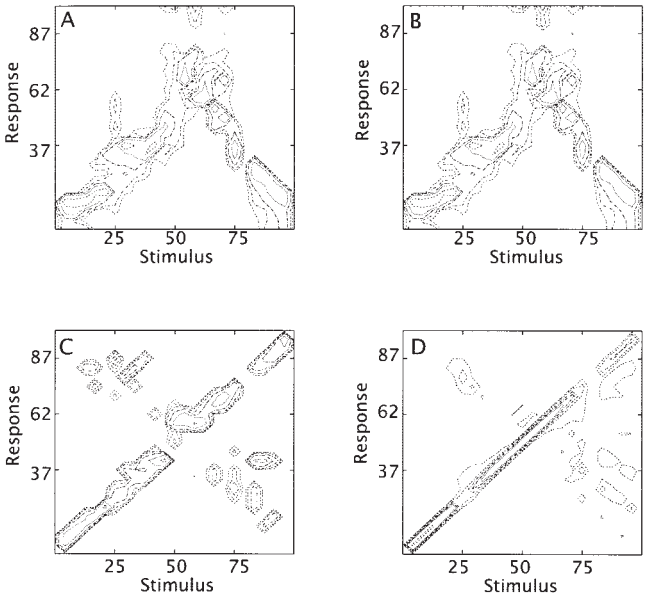

Figure 11. The best-fitting predictions of EXAM (extrapolation–association model) and POLE (population of linear experts model) to the bimodal response distribution of Participant 4 from Figure 10. A: EXAM with weights initialized to 0 at the outset. B: EXAM with preloading of weights. C: The participant's responses replotted as a frequency-contour plot for comparison. EXAM cannot produce bimodality in the transfer region. D: POLE's predicted distribution of responses.

previous experiment. The 2 (type of stimuli) × 4 (training block) within-participants ANOVA again revealed main effects of type of stimuli, $F(1, 44) = 243.38$, $MSE = 131.89$, $p < .01$, and training block, $F(3, 132) = 16.67$, $MSE = 16.93$, $p < .01$. Unlike the previous study, the interaction between both variables was also highly significant, $F(3, 132) = 5.66$, $MSE = 17.37$, $p < .01$. The interaction likely reflected the greater improvement with training for the exceptional stimuli than for the positive set.

As in the previous experiment, the responses of participants to the positive items were used to predict, via linear regression, the responses to the three exception items if their exception status had been ignored. Table 4 shows the observed mean responses to the exception items along with the responses predicted by the regression. As in the previous experiment, responses to the exception items changed during training, away from the response predicted by the positive stimuli and toward the correct magnitudes.

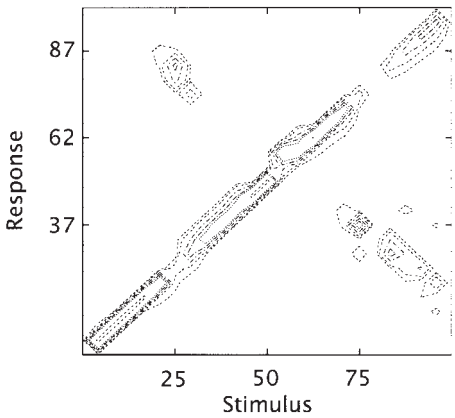

Figure 12. The predicted distribution of response frequencies for the last block of trials in Experiment 2, as the result of the single best-fitting set of parameters for POLE (population of linear experts model; see Table 1).

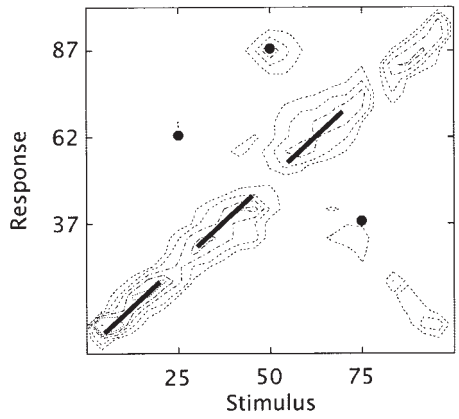

Figure 13. Observed frequency distribution (contour plot) of responses for the last block (including both training and transfer items) and training stimuli (thick lines and solid circles) for Experiment 3. Results are aggregated across all 45 participants. Note the bimodality in the transfer region ($x > 80$).

Transfer Performance

Figure 13 shows all responses made during the last block of the experiment. The most striking aspect of the data is that people again appeared to learn several conflicting functions simultaneously. This occurred even though the three exceptional stimuli did not lie on a single simple function. There again appeared to be considerable multimodality during transfer, which was statistically explored at the aggregate and individual participants' level.

Aggregate analysis. A total of 1,251 transfer responses were available for the aggregate analysis (9 missing observations). The cumulative frequency distribution of the residualized responses (in 103 bins of unit width) gave rise to a highly significant dip statistic (0.252, $p < .01$), once again confirming the presence of bimodality.

Individual-participants analysis. As in the previous study, separate dip tests were applied to the 28 transfer responses of each participant. Of those 45 tests, 23 were found to be significant at the .05 level, confirming that a sizeable proportion of participants alternated between different response functions. The nature of the observed bimodality is illustrated in Figure 14, which shows responses of 4 representative participants. Figures 14A and 14B show the responses of participants who exhibited least bimodality, whereas the responses in Figures 14C and 14D are from the participants who exhibited the greatest extent of bimodality.

Theoretical Analysis

Fit of EXAM

EXAM was fit to the data from a single participant (Participant 4 from Figure 14) who exhibited strong bimodality. The results

Table 4
The Prediction Equation and the Difference Between Responses to Exception Items in Experiment 3

Block	Equation	Exception 1	Exception 2	Exception 3
1	$y = 12.51 + .77x$	33.1 (31.8)	55.1 (51.0)	65.0 (70.3)
2	$y = 7.51 + .86x$	31.8 (29.0)	64.4 (50.5)	58.0 (72.0)
3	$y = 5.57 + .90x$	34.0 (28.1)	68.9 (50.6)	59.7 (73.1)
4	$y = 5.99 + .89x$	34.0 (28.3)	69.8 (50.5)	60.3 (72.7)
Training values	$y = x$	62 (25)	87 (50)	37 (75)

Note. Values in parentheses are predicted values based on the regression equation.

resemble those obtained for Experiment 2. As shown in Figure 15A, the two-parameter version failed to provide any extrapolations along the positive function ($B = 180.33$, $\gamma = 0.03$, $\eta = 0.07$) and also exhibited no bimodality for any of the trained magnitudes. The model furthermore consistently underpredicted response magnitudes for the outlying training stimulus at $x = 25$.

The performance of the three-parameter version, with preinitialized weights, was little better. As shown in Figure 15B, EXAM's predictions for the extrapolation region were all unimodal and included implausible negative magnitudes for all test stimuli with $x > 88$ ($B = 171.58$, $\gamma = 0.05$, $\eta = 0.08$, $\beta = .49$).

Fit of POLE

POLE was fit to each participant individually, resulting in a mean fit of $B = 33.71$ ($SD = 7.29$). Figure 15D shows POLE's predictions (with best-fitting parameters $c = 35.00$, $\eta_s = 22.19$, $\lambda_w = 0.70$, $\lambda_b = 0.30$, $\omega = 1.81$, $\epsilon = 24.80$, and $B = 71.39$) for the same participant that EXAM failed to model. The empirically observed bimodality for training and transfer stimuli is clearly captured in these predictions.

We again constrained POLE to fit the data with a single set of parameters. The best-fitting six parameters (listed in Table 1)

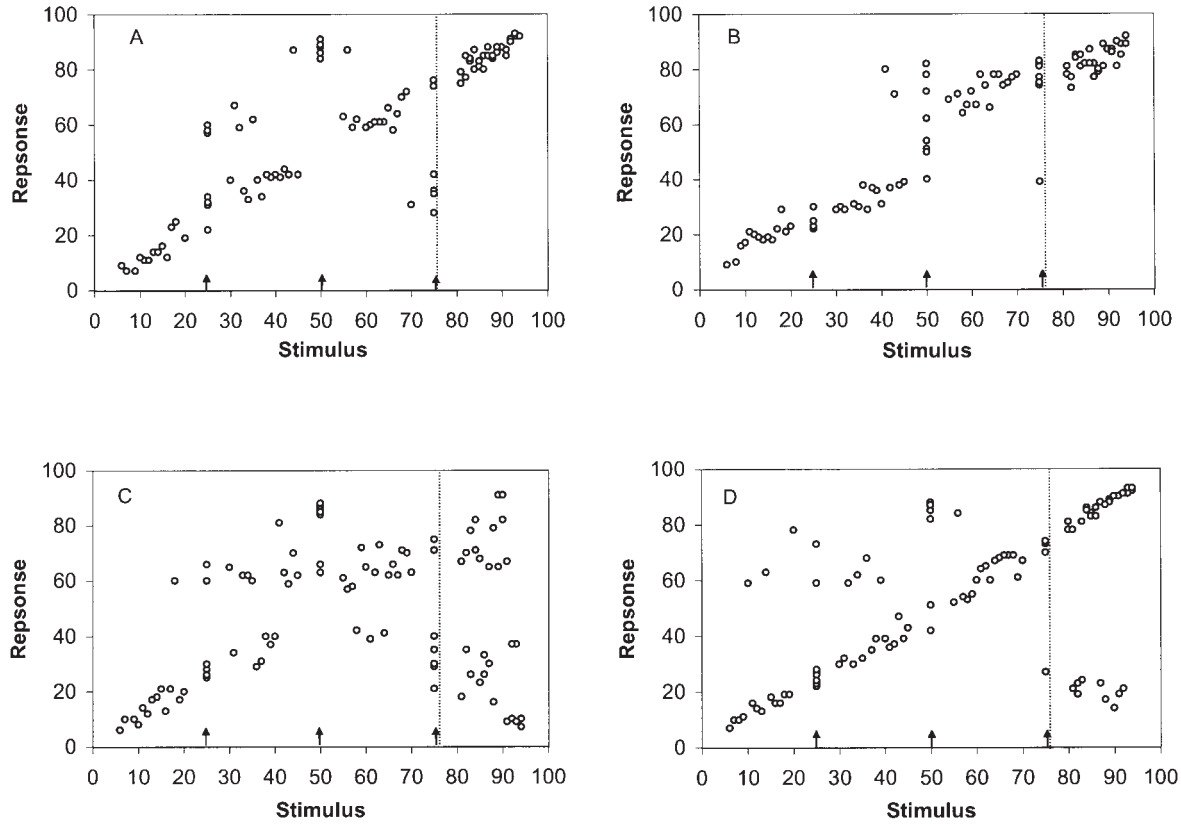

Figure 14. Responses of 4 participants from Experiment 3, showing the diversity of response patterns to the final block of stimuli. The dashed vertical line in each panel marks the boundary of the training stimuli; items to the right of the line were presented without feedback only. The three small arrows on each *x*-axis mark the locations of the high-frequency exception items. Dip scores are 0.040, 0.042, 0.074, and 0.255 for Participants 1–4 (shown in A–D), respectively.

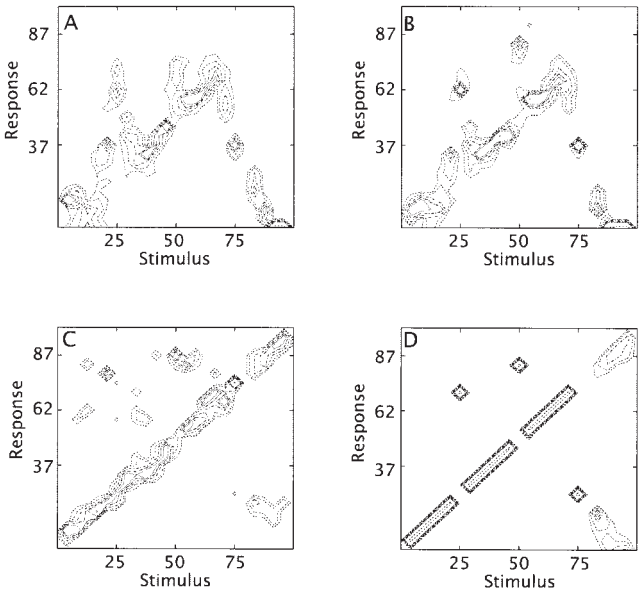

Figure 15. The best-fitting predictions of EXAM (extrapolation–association model) and POLE (population of linear experts model) to the bimodal response distribution of Participant 4 from Figure 14. A: EXAM with weights initialized to 0 at the outset. B: EXAM with preloading. C: The participant's responses replotted as a frequency-contour graph for comparison. EXAM cannot produce bimodality in either the transfer or training regions. D: POLE's predictions.

produced a mean fit of $B = 34.15$ ($SD = 7.38$), an increase of only 1%, or 6% of a standard deviation. Figure 16 shows the predicted frequency distribution for the final block of the experiment, which is again clearly multimodal. As in Experiments 1 and 2, the data from Experiment 3 strongly support POLE's core prediction of competitive selection of simple response functions when learning a complex function concept.

GENERAL DISCUSSION

This article has been organized around the central theme of knowledge partitioning. We presented a computational model of function learning, POLE, which assumed that all function learning relies on the splicing of partitioned knowledge. POLE accounted for the knowledge partitioning seen in previous experiments (Lewandowsky et al., 2002) and also captured benchmark results in function learning. In addition, the model's predictions concerning multimodality within the responses of individual participants were confirmed in three experiments.

We now take up these points in reverse order. We first analyze our empirical contribution, we then explore POLE further (in particular its relation to empirical and theoretical precursors), and we conclude with a discussion of the relationship between our results and the general framework of knowledge partitioning.

Empirical Contribution

Our three experiments converged on a single strong conclusion: In situations of uncertainty people select very different responses to the same stimulus on different occasions. There was little if any evidence that people ever averaged or blended competing responses. Instead, multimodality was present in all three experiments notwithstanding considerable variability among the to-be-learned functions. Experiment 1 presented participants with two linear segments differing only in their intercept, Experiment 2 presented one dominant increasing function and three outliers that together formed a decreasing linear function, and Experiment 3 presented the same increasing function but used outliers that were not linearly related. The generality of multimodality is further supported by the fact that in Experiments 2 and 3, participants exhibited multimodality during training as well as on novel transfer items.

Multimodality similar to that observed here was also reported by Kruschke (2001a) in a study that sought evidence for the presence of blocking and highlighting (i.e., the inverse base rate effect) in function learning. Kruschke showed that both blocking and highlighting occur in function learning much like they do in categorization (e.g., Kruschke, 1996). Kruschke (2001a) additionally discovered that responses were bimodal. That is, people selected one or the other of two functions when responding, on the basis of two competing cues, without any evidence of averaging of the implied responses.

The observed multimodality underscores the diagnosticity of function learning in general, particularly in comparison to category learning. At first glance, the differences between categorization and function learning appear to be fairly small: Whereas responses in category learning are nominal, involving arbitrary labels without any implied order or numeric magnitude, responses in function learning are explicitly ordered along a magnitude axis. In all other respects, the learning of categorical and function concepts share much in common; there are predictors whose relationship to the responses must be learned from a set of training instances; after ample training, people are asked to generalize or extrapolate their knowledge to new test items; and people are remarkably adept at such extrapolation and generalization. Accordingly, the EXAM (DeLosh et al., 1997) theory of function learning explicitly acknowledges a theory of categorization (ALCOVE; Kruschke, 1992) among its conceptual antecedents, just as POLE does (ATRIUM; Erickson & Kruschke, 1998).

However, the surface similarity and theoretical linkage between categorization and function learning obscures a significant differ-

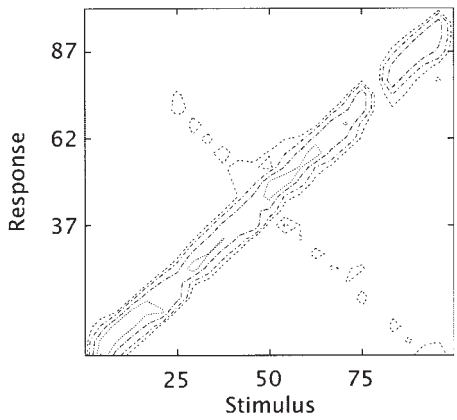

Figure 16. The predicted distribution of response frequencies for the last block of trials in Experiment 3, as the result of the single best-fitting set of parameters for POLE (population of linear experts model; see Table 1).

ence in potential diagnosticity (cf. Kruschke, 2001a). The multimodality revealed in the present experiments could not have been observed in categorization because the nominal response scale lacks potential intermediate response alternatives. There is no sense in which we can inspect a response and see whether it is the average of two nominal responses: It must always be one or the other. Because multimodality is a sign of knowledge partitioning, it follows that the partitioning of knowledge in the present experiments may also routinely occur in categorization, although it may not be readily identifiable by behavioral means. In support, Yang and Lewandowsky (in press) showed that a mixture-of-experts model (ATRIUM) accounted for learning of a complex categorization task by partitioning the underlying representation into multiple independent modules.

The occurrence of bimodality also serves to differentiate function-concept learning from results of loosely related motor-sensory adaptation paradigms. For example, Vetter and Wolpert (2000, their "rotation experiment") trained people to reach out and touch a single target under one of two feedback conditions: veridical and rotated (so that the finger and target did not appear to the participant to be where they really were). Participants initially learned to move accurately under each condition separately. At a subsequent test, limited information was provided about the condition of a given movement. This uncertainty caused participants to execute the average of the two movements, rather than alternating between them (see also Ghahramani & Wolpert, 1997; Scheidt, Dingwell, & Mussa-Ivaldi, 2001). The absence of bimodality under those circumstances may suggest that different learning systems underlie motor and concept learning, or it may reflect differences in the experimental paradigms. It remains for future research to elaborate on a possible empirical linkage between function learning and motor learning.

Theoretical Contribution: POLE, a New Theory of Function Learning

Like category learning, function learning results in the formation of a new concept. The theory proposed in this article identifies several mechanisms thought to underlie this concept formation: error-driven learning, instance-based representations, prior expectations, dimensional attention, and competition between existing candidate functions. These last two features distinguish this theory from its predecessors and are essential to its success.

POLE accurately predicted both the presence and absence of knowledge partitioning under different training conditions reported by Lewandowsky et al. (2002). Dimensional attention and competition between experts were central to these predictions. When function segments were correlated with different contexts, error could be rapidly reduced by shifting attention to the context dimension and dividing the to-be-learned function into approximately linear components for each context. When context was random, a common nonlinear function had to be learned, which POLE achieved by carefully balancing the weights from many similar stimuli to switch between competing experts as the stimulus went through small changes.

Principles of Function Learning and POLE

POLE's underlying mechanisms were also sufficient to account for 8 of 10 principles of function learning identified by Busemeyer et al. (1997). We now turn to a more speculative discussion of the last two principles.

Principle 9 states that functions are easier to learn if the cue labels are congruent with expectations (e.g., severity of car accidents would be expected to increase with driving speed). A plausible interpretation of the role of cue labels is that they determine the nature of people's a priori biases, expressed within POLE by Equation 4 as a positive unit slope (i.e., $M = 1$). Thus, labels that indicate an inverse relationship (such as between fatigue and hours of sleep) may simply change the sign of M without any further effect within the architecture. For labels that indicate a more complex relationship, such as between date and fullness of the moon, we offer two suggestions. The first is that people construct a derived stimulus representation. In the lunar example, if the date were presented as day of the year (1–365), people might take the modulus of 28 and so transform the complex cyclic function into a simpler function with only a single cycle. As suggested by our analysis of learning different function types, POLE with its standard preferred slope can learn such single-cycle functions quite well, at least within the limits of the training stimuli. The second suggestion is that people, if absolutely necessary, may use nonlinear experts. We thus remain open to the possibility that the linear experts in POLE may sometimes be replaced or augmented by nonlinear ones, although we do not explore this possibility here.

Principle 10 captures the fact that learning is accelerated if training stimuli are presented in systematic order, for example from smallest to largest magnitudes. We identify three ways in which this effect might be captured by POLE. First, neighboring instances in POLE are activated jointly, to an extent determined by their specificity (c in Equation 2). Specificity determines how learning generalizes between stimuli, and systematic presentation may inform participants about the optimal value for c more rapidly than random presentations. The second reason has to do with short-term memory, a process not explicitly considered in most models of category or function learning. If participants can remember which stimuli were presented on the last few trials and what responses were correct for those items, then the rate of learning can be improved through rehearsal strategies. Finally, in line with our conjecture about Principle 9, systematic presentation may provide a better environment for the extraction of higher order stimulus properties that could, in turn, simplify the task.

POLE's Theoretical Neighborhood

Although POLE is a new theory, it has a close connection to existing approaches to concept formation. Within the field of function learning, it shares with EXAM an instance-based approach, and it shares with Koh and Meyer's (1991) adaptive regression model the use of parametric functions (albeit only linear ones) and an initial bias toward certain functions. It is important to note that unlike Koh and Meyer's model but like EXAM, POLE uses the error from each trial as the impetus for changing internal parameters (weights and strengths). Koh and Meyer's model instead uses a set of internal parameters that is optimal for the entire set of presented trials (subject to certain constraints). A new technique in adaptive regression, using an instance-based mixture of experts (Schaal & Atkeson, 1998), shares many of POLE's properties and might provide an interesting contrast were it to be developed into a psychological model of human learning.

POLE is also closely related to models of category learning that are based on the mixture-of-experts framework. Most notable among these is ATRIUM (Erickson & Kruschke, 1998, 2002; Kruschke & Erickson, 1994). In ATRIUM, instance-specific weights are combined with nonspecific weights to determine the relative strength of different experts. The experts in ATRIUM are single-dimensional response rules, which are analogous to the experts in POLE. ATRIUM additionally uses collective instance memory as an expert in its own right, which is an aspect of the model that has no analogue in POLE. Moreover, the nonspecific weights in ATRIUM are tied to stimulus dimensions, whereas in POLE each expert has a single bias that is not tied to a particular stimulus dimension (e.g., context vs. magnitude). Finally, and most important, ATRIUM computes predictions by blending together the output from various experts, rather than probabilistically choosing a single expert to govern responding on each trial. Empirically, blending of experts can be differentiated from probabilistic choice in a function-learning paradigm (through multimodality) but, for the reasons discussed earlier, not in category learning.

Finally, we return to the suggested connection between function learning and perceptual-motor learning (Rosenbaum et al., 2001). In this vein, we note that several recent theoretical developments in motor learning bear some resemblance to POLE. People's ability to learn complex visuomotor mappings (Shadmehr & Mussa-Ivaldi, 1994) has been attributed to the acquisition of new "internal models" (Conditt et al., 1997) that have been conceived to be "experts" much like in POLE, but with complex nonlinear properties (Ghahramani & Wolpert, 1997; Haruno et al., 2001). The MOSAIC model (Haruno et al., 2001) uses multiple experts (one for each stimulus) to explain the ability of people to interpolate between learned stimuli and to switch between experts when the stimulus changes. MOSAIC differs from POLE in many ways; it is not constrained to linear experts, it blends rather than chooses experts, it does not use instance information to weight experts prior to mixing them, and it has no conception of dimensional attention.

Limitations and Possible Extensions

Throughout this article, we compared POLE with EXAM, the hitherto most successful model of function learning. This comparison revealed that POLE fits our individual-participants data considerably better than does EXAM, although at the expense of having three more parameters. It is therefore appropriate to ask whether these additional parameters yield a sufficiently large increase in explanatory power. Visual analysis of the behavior of the two models (i.e., fits to Experiments 2 and 3) clarifies that there are no parameter values that can lead even the extended version of EXAM to predict a substantial number of bimodal responses and still retain any accuracy on the training stimuli. In contrast, POLE makes this prediction by virtue of its basic architecture, across a range (but not the whole range) of parameter values. The additional three parameters in POLE yield a qualitative increase in explanatory power; a quantitative comparison is unnecessary to distinguish the models.

Despite its additional parameters, the current instantiation of POLE rests on a few parsimonious assumptions: Experts constitute a population that is linear and one-dimensional with fixed slopes and intercepts. Further, selection of experts depends only on associations with stored instances and nonspecific biases. These assumptions entail several limitations.

The first limitation is a practical one; although the theory makes reference to a population of experts, in the present simulations that population was restricted to 64 candidate functions. However, anecdotal simulations suggested that the precise number does not matter qualitatively, so long as the experts span the function space with reasonable density.

The second limitation is the functional form, and flexibility, of the experts. Appendix B shows that fixed slopes and intercepts is not a necessary constraint, and indeed, Schaal and Atkeson's (1998) algorithm combines learning of expert's internal parameters with instance-specific changes in sensitivity. To date, we have not explored situations in which it may be necessary for people to adjust the slopes and intercepts of their experts. Similarly, the decision to include only linear functions could be relaxed because what counts as simple may be a task-dependent decision by the individual (e.g., in some contexts, cyclical functions may be simple).

The third potential limitation of the model is the nature of the expert selection process. As suggested earlier, more complex stimulus representations, such as recoding of cyclic functions to remove periodicity, may be required in certain situations. There is good reason to believe that people are able to adapt their representation of stimuli to suit the concept being learned (Goldstone, Lippa, & Shiffrin, 2001; Schyns, Goldstone, & Thibaut, 1998), but these processes are not currently formalized in POLE.

Knowledge Partitioning

POLE represents a specific computational instantiation of knowledge partitioning in function learning and assumes that partitioning underlies all function learning. However, the present results and modeling have wider theoretical implications.

Related Previous Findings

Research on mental arithmetic has also uncovered a prolonged coexistence of alternative strategies and forms of knowledge that is reminiscent of partitioning in function learning. Reder and Ritter (1992) and Schunn, Reder, Nhouyvanisvong, Richards, and Stoffolino (1997) presented participants repeatedly with two-digit $\times$ two-digit multiplication problems (e.g., 43×19). Prior to answering a problem, participants had to rapidly indicate whether they could retrieve the correct answer from memory (which they then had to do within a short time) or whether they would need to compute the answer (in which case extra time was allotted). Across repeated presentations of a given problem, people were found to switch strategies not just once but between two and three times, and switches were separated by up to 50% of all learning trials (reported in Delaney, Reder, Staszewski, & Ritter, 1998), suggesting that both forms of knowledge—retrieval and computation—continued to coexist throughout the training sequence. Prolonged coexistence of alternative arithmetic knowledge has also been observed at a much larger time scale, across grades in primary school (e.g., Shrager & Siegler, 1998; Siegler, 1987).

These findings have been echoed at a theoretical level by Lovett and Schunn (1999) in a general model of choice during problem solving, known as RCCL (pronounced "ReCyCLe"). Central to RCCL is the availability of a set of alternative strategies to solve

a common problem. RCCL posits that people choose a strategy for each trial on the basis of the past success of the available alternatives. In support, Lovett and Schunn reported two experiments in which people repeatedly switched between strategies when no one strategy was particularly successful; conversely, people tended to persist with a successful strategy.

In contrast to the present results and the general knowledge partitioning framework, none of the preceding studies—and other context effects surveyed at the outset—showed that these coexisting strategies could engender contradictory behaviors. That is, the solution to 19×23 can be obtained by direct memory retrieval or by computation, and the two strategies may entail different completion times (e.g., Delaney et al., 1998; Schunn et al., 1997), but they both lead to the same correct answer. Indeed, there is every reason to expect that people would rapidly abandon any strategy that gives rise to errors that are avoided by use of an alternative (Lovett & Schunn, 1999). Thus, the contradictory multimodality of our results, and others within the knowledge partitioning framework, remains a unique empirical contribution.

Continued Partitioning Versus Incremental Transition

By itself, the idea that different forms of knowledge can support performance on the same task is not new. For example, the instance theory of automaticity (e.g., Logan, 1988) postulates that acquisition of a cognitive skill consists of the transition from an initial slow algorithm to the fast retrieval of memorized solutions to previously encountered problems (see also knowledge compilation, Anderson & Fincham, 1994, and component power law, Rickard, 1997).

Common to these views is the unidirectionality and finality of the transition between different forms of knowledge: It is assumed that people gradually abandon their initial approach to the task in favor of a maturing alternative that eventually provides the ongoing and sole basis for task performance. Knowledge partitioning, by contrast, holds that different parcels of knowledge can simultaneously emerge during skill acquisition and, importantly, continue to coexist and compete during task performance.

This longevity of partitioning was observed in the present experiments, the function learning experiments reported by Lewandowsky et al. (2002), and the category learning studies by Yang and Lewandowsky (2003, in press). Indeed, the contradictory nature of expert knowledge reported by Lewandowsky and Kirsner (2000) and Schliemann and Carraher (1993) is compatible with the basic assumption of knowledge partitioning that independent parcels of knowledge may coexist in perpetuity. POLE's account of partitioning assumes that it occurs to achieve rapid error reduction. Indeed, partitioning may be a prerequisite to automaticity: Automaticity relies on consistent stimulus–response mappings (Schneider & Shiffrin, 1977), and partitioning of a complex problem space may be the only way to achieve such mappings in practice.

CONCLUSION

Function learning provides a uniquely powerful tool for the examination of concept learning. Three experiments showed that people produced bimodal responses in situations of uncertainty; in a categorization experiment, the partitioning underlying this multimodality would have gone unnoticed. Inspired by recent results in diverse areas, we presented a model of function learning that simplifies difficult problems through the partitioning of the stimulus space into simpler regions. The model, POLE, provided a good quantitative account of knowledge partitioning. The model also predicted basic properties of function learning and the quantitative results of specific studies. POLE proved to be the only extant model of function learning able to predict the bimodality observed in our experiments because of its unique combination of psychological principles, including the learned probabilistic, exemplar-specific, selection of competing, and simple response alternatives.

References

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1322–1340.

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

Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research, 41,* 449–461.

Bédard, J., & Chi, M. T. H. (1992). Expertise. *Current Directions in Psychological Science, 1,* 135–139.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology, 89,* 417–429.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3,* 21–29.

Conditt, M., Gandolfo, F., & Mussa-Ivaldi, F. (1997). The motor system does not learn the dynamics of the arm by rote memorization of past experience. *Journal of Neurophysiology, 78,* 554–560.

Crawford, J. D., & Guitton, D. (1997). Primate head-free saccade generator implements a desired (post-VOR) eye position command by anticipating intended head motion. *Journal of Neurophysiology, 78,* 2811–2816.

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

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

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

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review, 9,* 160–168.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 1–50). Hillsdale, NJ: Erlbaum.

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

Ghahramani, Z., & Wolpert, D. (1997, March 27). Modular decomposition in visuomotor learning. *Nature, 386,* 392–395.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Glaser, R. (1996). Changing the agency for learning: Acquiring expert performance. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 303–311). Hillsdale, NJ: Erlbaum.

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted positions: Implications for the theory of expertise. *Memory & Cognition, 24,* 493–503.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78,* 27–43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130,* 116–139.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics, 13,* 70–84.

Haruno, M., Wolpert, D., & Kawato, M. (2001). MOSAIC model for sensorimotor learning and control. *Neural Computation, 13,* 2201–2220.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

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

Jacobs, R. A., & Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. *Vision Research, 39,* 4062–4075.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

Koh, K. (1993). Induction of combination rules in two-dimensional function learning. *Memory & Cognition, 21,* 573–590.

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.

Kravitz, J., & Yaffe, F. (1972). Conditioned adaptation to prismatic displacement with a tone as the conditioned stimulus. *Perception & Psychophysics, 12,* 305–308.

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

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

Kruschke, J. K. (2001a, July). *Cue competition in function learning.* Talk presented at the 3rd International Conference on Memory, Valencia, Spain.

Kruschke, J. K. (2001b). Toward a unified model of attention in association learning. *Journal of Mathematical Psychology, 45,* 812–863.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1083–1119.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1666–1684.

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..

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & Cognition, 28,* 295–305.

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

Lovett, M. C., & Schunn, C. D. (1999). Task representations strategy variability and base-rate neglect. *Journal of Experimental Psychology: General, 128,* 107–130.

Martin, T., Keating, J., Goodkin, H., & Bastian, A. (1996). Throwing while looking through prisms: II. Specificity and storage of multiple gaze-throw calibrations. *Brain, 119,* 1199–1211.

Mellers, B. (1986). Test of a distributional theory of intuitive numerical prediction. *Organizational Behavior and Human Decision Processes, 38,* 279–294.

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

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics.* Cambridge, England: Cambridge University Press.

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

Reed, S. K., & Evans, A. C. (1987). Learning functional relations: A theoretical and instructional analysis. *Journal of Experimental Psychology: General, 116,* 106–118.

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

Roe, R., Barkan, R., & Busemeyer, J. (2001, July). *Extrapolation in multiple-cue function learning: Critical tests of associative versus rule based models.* Paper presented at the 3rd International Conference on Memory, Valencia, Spain.

Rosenbaum, D., Carlson, R., & Gilmore, R. (2001). Acquisition of intellectual and motor skills. *Annual Review of Psychology, 52,* 453–470.

Schaal, S., & Atkeson, C. (1998). Constructive incremental learning from only local information. *Neural Computation, 10,* 2047–2084.

Scheidt, R., Dingwell, J., & Mussa-Ivaldi, F. (2001). Learning to move amid uncertainty. *Journal of Neurophysiology, 86,* 971–985.

Schliemann, A. D., & Carraher, D. W. (1993). Proportional reasoning in and out of school. In P. Light & G. Butterworth (Eds.), *Context and cognition: Ways of learning and knowing* (pp. 47–73). Hillsdale, NJ: Erlbaum.

Schneider, W., & Shiffrin, R. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stoffolino, P. J. (1997). To calculate or not calculate: A source activation confusion model of problem-familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 3–29.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–54.

Shadmehr, R., & Mussa-Ivaldi, F. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience, 14,* 3208–3224.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9,* 405–410.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Turvey, M. T. (1996). Dynamic touch. *American Psychologist, 51,* 1134–1152.

Vetter, P., & Wolpert, D. (2000). Context estimation for sensorimotor control. *Journal of Neurophysiology, 84,* 1026–1034.

Welch, R., Bridgeman, B., Anand, S., & Browman, K. (1993). Alternating prism exposure causes dual adaptation and generalization to a novel displacement. *Perception & Psychophysics, 54,* 195–204.

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 663–679.

Yang, L.-X., & Lewandowsky, S. (in press). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Appendix A

Implementation Details of EXAM

This appendix summarizes the description of EXAM provided by DeLosh et al. (1997) along with two novel extensions: one necessary to predict a probabilistic set of responses for each stimulus and the other needed to capture preexperimental biases in responses. Because this description follows that of DeLosh et al., we have tried to keep our notation as close to theirs as possible. This necessarily introduces differences with the notation used to describe POLE in the body of the article.

Architecture and Learning Rule

EXAM can be considered as a connectionist network, with a large number of instance nodes representing both stimuli and responses. Each unique stimulus–response pair presented during training is represented by a pair of input and output nodes at the corresponding locations (denoted X and Y, respectively) on the real number line. All M of these input and output nodes are fully interconnected by a single layer of learned weights.

Presentation of a stimulus of magnitude X activates each input node i according to a Gaussian distance function:

$$a_i(X) = \exp -\gamma(X - X_i)^2, \tag{A1}$$

where γ is a scaling parameter that determines the steepness with which activation declines as a function of the distance between the presented stimulus magnitude, X, and the location of the node, X_i. Each training stimulus maximally activates the input node closest to its location on the number line but also contributes to the activation of neighboring nodes.

Activation passes through the layer of weights to the output nodes, where it is summed to form the activation of an output unit j:

$$o_j(X) = \sum_{i=1}^{M} w_{ji} a_i(X), \tag{A2}$$

where w_{ji} represents the weight between input unit i and output unit j. In the simulations reported by DeLosh et al. (1997), all weights were initialized to 0. Unless noted otherwise, we follow the same practice here.

Learning takes place on the basis of a feedback signal that considers the difference between obtained and desired activations of all output units. Regardless of how weights are initialized, they were adjusted during training using conventional error-driven learning:

$$w_{ji}^{\text{new}} = w_{ji} + \eta \Delta_{ji}, \tag{A3}$$

where η represents a learning rate parameter and Δ_{ji} is the weight update given by

$$\Delta_{ji} = (f_j(Z) - o_j(X)) a_i(X), \tag{A4}$$

where $f_j(Z)$ is the feedback signal for output node j provided by the target magnitude Z. The feedback signal mirrored the activation function of the input nodes by including a Gaussian similarity gradient:

$$f_j(Z) = \exp - \gamma(Z - Y_j)^2. \tag{A5}$$

Thus, the feedback signal was maximal for the output unit at the location of the target magnitude (Z), but feedback generalized to neighboring nodes as well.

Response Generation: Interpolation and Extrapolation

When presented with a stimulus, a pure exemplar memory (e.g., EXAM as described so far) computes its output on the basis of the average response given to similar instances in memory. This simple response rule has been shown to be insufficient to account for people's performance in function learning tasks (Busemeyer et al., 1997). For this reason, EXAM departs from a pure instance-based representation by implementing a more complex response rule. When a test stimulus is presented, it is matched to a trained instance with a probability determined by the psychological similarity (i.e., magnitude difference) of the stored instances to the test stimulus. The matched instance is then used as a cue to retrieve three previously learned response magnitudes: that associated with the matched instance and those of the two immediate neighbors to either side. These three response magnitudes, in turn, are used to produce a local slope estimate for interpolation (or extrapolation, if the stimulus is outside the trained range).

In practice, EXAM generates overt responses by a sequence of probabilistic matching steps that implement this response rule. First, the presented test stimulus X is matched to each input node i with a probability given by Luce's choice rule:

$$P(i|X) = \frac{a_i(X)}{\sum_{k=1}^{M} a_k(X)}. \tag{A6}$$

Given that X is matched to node i, the next step involves retrieval of mean output values for that node plus its immediate neighbors to either side (i.e., $i - 1$ and $i + 1$; i is used if it is either the lowest or highest stimulus value encountered during training). These mean output values, denoted $m(X_i)$, $m(X_{i-1})$, and $m(X_{i+1})$, are used for a local slope estimate which in turn allows extrapolation (or interpolation) from those trained response magnitudes to the test stimulus X. Mean output values are given by

$$m(X) = \sum_{j=1}^{M} Y_j P(j|X_i), \tag{A7}$$

where $P(j|X_i)$ is provided by Luce's choice rule:

$$P(j|X) = \frac{o_j(X)}{\sum_{k=1}^{M} o_k(X)}, \tag{A8}$$

with $o_k(X)$ as defined by Equation A2. Hence, the activation profile across all output units maps into the probabilities with which all potential response magnitudes (i.e., the Y_js) contribute to the mean output in response to a particular stimulus X_i.

The mean output values given by Equation A7 are then incorporated into a response:

$$E(Y|X_i) = m(X_i) + \frac{m(X_i + 1) - m(X_i - 1)}{(X_i + 1) - (X_i - 1)} (X - X_i). \tag{A9}$$

The right-hand side of Equation A9 contains two distinct components: The first, $m(X_i)$, represents the output value associated with the training item matched to the test stimulus. The remaining component provides linear interpolation from that training item to the test stimulus using a slope estimate provided by the output values retrieved for the neighboring instances.

Finally, the model's overt response to X is given by summing the responses provided by Equation A9 across all input units to which the test stimulus might be matched:

$$E(Y|X) = \sum_{i=1}^{M} P(i|X) E(Y|X_i). \tag{A10}$$

(*Appendixes continue*)

The mean response provided by Equation A10 constitutes EXAM's predictions. All fits reported by DeLosh et al. (1997) and some of the fits reported in this article minimized the root-mean-squared deviations between these predicted mean responses and the empirically obtained response magnitudes.

Modifications to EXAM

We made two modifications to EXAM to improve its chance of fitting our data. First, EXAM does not explicitly predict a distribution of response magnitudes. However, it is possible to consider the terms entering into the summation in Equation A10 as components of a distribution of predicted responses. Thus, each possible response magnitude $E(Y|X_i)$ has some probability $P(i|X)$ of occurrence that can be compared to the empirically obtained values.

Second, for some of the simulations reported in this article, we initialized the weights that directly connected input and output nodes (i.e., all w_{ii}s for $i = 1, 2, \ldots, M$) to some constant value specified by the (third) free parameter β. This captured people's known (e.g., Busemeyer et al., 1997) expectation that all functions are linearly positive.

Appendix B

Derivation of POLE's Learning Rule

Shifting Response Strengths Between Experts

To begin the derivation of learning rules, we recall from Equation 8 that error is defined by

$$E_{\text{Mix}} = \sum_k S_k E_k. \tag{B1}$$

For the distribution of expert strengths to be shifted so as to reduce error, the change must be in the direction of the negative gradient:

$$\begin{aligned}\Delta s_K &= -\eta_s \frac{\partial E_{\text{Mix}}}{\partial s_K} \\ &= -\eta_s \sum_k E_k \frac{\partial S_k}{\partial s_K} \\ &= -\eta_s \sum_k E_k(\kappa_{kK} - S_k)/\sum_j s_j \\ &= \eta_s(E_{\text{Mix}} - E_K)/\sum_j s_j . \end{aligned} \tag{B2}$$

Learning Within Each Expert

There is nothing to prevent the intercept and slope parameters within each expert from being learned, although in the current formulation they remain constant. We derive the rules for the linear experts here for completeness:

$$\Delta m_k = -\lambda_m \frac{\partial E_{\text{Mix}}}{\partial m_k} = \lambda_m S_k(y - \hat{y}_k)x_1, \tag{B3}$$

which would adjust the slope of each expert, and

$$\Delta b_k = -\lambda_b \frac{\partial E_{\text{Mix}}}{\partial b_k} = \lambda_b S_k(y - \hat{y}_k), \tag{B4}$$

which would adjust the intercept to reduce error.

Learning of Strength Node Weights

The weight update mechanism defined by Equation 10 derives from the fact that error can be conveniently defined at the level of the strength nodes by taking the difference between the shifted strengths and the initial strengths:

$$E_{\text{Strength}} = \frac{1}{2}\sum_k (s_k^{\text{shift}} - s_k^{\text{init}})^2. \tag{B5}$$

This is the same as the error at the level of the output nodes, because the shifted strengths were driven from their initial values by the output error. The local computation of this error improves the approximation to the true error gradient, because the strengths have shifted iteratively.

From this definition of error, we can derive the gradient of error with respect to the weights from exemplar nodes to strength nodes and adjust the weights accordingly:

$$\begin{aligned}\Delta w_{kj} &= -\lambda_w \frac{\partial E_{\text{Strength}}}{\partial w_{kj}} \\ &= \lambda_w(s_k^{\text{shift}} - s_k^{\text{init}})s_k^{\text{init}}a_j^{\text{Inst}}. \end{aligned} \tag{B6}$$

Similarly, the gradient of the error with respect to bias weights is given by

$$\begin{aligned}\Delta \omega_{k0} &= -\lambda_b \frac{\partial E_{\text{gain}}}{\partial \omega_{k0}} \\ &= \lambda_b(g_k^{\text{shift}} - g_k^{\text{init}})\underbrace{g_k^{\text{init}}/\omega_{k0}}_{\exp(\sum_j w_{kj}a_j^{\text{in}})}, \end{aligned} \tag{B7}$$

with the expression in the underbrace being the preferred computational form because the bias weight may happen to be 0. In the simulations, if learning drives the bias weight to a negative value, it is clipped at 0.

Learning Dimensional Attention

The propagation of error down to the input dimension attractions is a more complicated calculation. As an initial step, it is useful to determine the derivative of the dimensional attention with respect to the attraction. Recall that Equation 1 specifies this relation. Differentiation yields

$$\begin{aligned}\frac{\partial \alpha_i}{\partial \aleph_I} &= \left[\sum_j \exp(\aleph_j)\frac{\partial \exp(\aleph_i)}{\partial \gamma_I} - \exp(\aleph_i)\frac{\partial \sum_j \exp(\aleph_j)}{\partial \aleph_I}\right] \Bigg/ \\ &\quad [\sum_j \exp(\aleph_j)]^2 \\ &= [\sum_j \exp(\aleph_j)\exp(\aleph_I)\kappa_{iI} - \exp(\aleph_i)\exp(\aleph_I)]/[\sum_j \exp(\aleph_j)]^2 \\ &= \alpha_I\kappa_{iI} - \alpha_i\alpha_I \\ &= (\kappa_{iI} - \alpha_i)\alpha_I. \end{aligned} \tag{B8}$$

Applying the chain rule, we can now derive the error at the expert strength nodes with respect to the dimension attractions. Boldface variables denote the vector with the corresponding variable as its components:

$$\begin{aligned}
\Delta\aleph_I &= -\lambda_{\text{dim}} \frac{\partial E_{\text{Strength}}}{\partial \aleph_I} \\
&= -\lambda_{\text{dim}} \frac{\partial E_{\text{Strength}}}{\partial \mathbf{s}^{\text{init}}} \frac{\partial \mathbf{s}^{\text{init}}}{\partial \mathbf{a}^{\text{Inst}}} \frac{\partial \mathbf{a}^{\text{Inst}}}{\partial \alpha} \frac{\partial \alpha}{\partial \aleph_I} \\
&= -\lambda_{\text{dim}}[\cdots - (s_k^{\text{shift}} - s_k^{\text{init}}) \cdots] \\
&\quad \times \begin{bmatrix} \vdots & \\ \cdots & s_k^{\text{init}} w_{kj} \quad \cdots \\ \vdots & \end{bmatrix} \\
&\quad \times \begin{bmatrix} \vdots & \\ \cdots & a_j^{\text{Inst}}(-c)|x_i - \mu_{ji}| \quad \cdots \\ \vdots & \end{bmatrix} \\
&\quad \times \begin{bmatrix} \vdots \\ \alpha_I(\kappa_{iI} - \alpha_i) \\ \vdots \end{bmatrix} \\
&= -\lambda_{\text{dim}} \sum_{k,j,i} (s_k^{\text{shift}} - s_k^{\text{init}}) s_k^{\text{init}} \\
&\quad \times w_{kj} a_j^{\text{Inst}} c|x_i - \mu_{ji}|(\kappa_{iI} - \alpha_i)\alpha_I. \qquad \text{(B9)}
\end{aligned}$$

Appendix C

Details of Model Fitting Procedures

To fit POLE and EXAM to the observed response frequency distributions in our three experiments, we discretized both X and Y into a number of bins. Thus, on each trial, a participant was presented with a stimulus in bin X and gave a response in bin Y. We found that 25 bins provided satisfactory stability in the optimization process.

The badness of fit on any given trial t was measured as

$$B_t = -\frac{1}{N}[\log P(Y|X) \sum_{\text{bins} \neq Y} \log(1 - P(\text{bin}|X))], \qquad \text{(C1)}$$

where $P(\text{bin}|X)$ is the predicted probability of producing a response in a given bin on that particular trial. This statistic must be computed trial by trial because of the nature of the experiment, which provides feedback after each trial and which therefore allows the predicted probabilities to change over trials.

Although badness of fit must be measured trial by trial, there is no reason why each trial must be weighted equally during optimization. In particular, one might be especially concerned with people's responses to transfer items, which are both novel and presented without feedback or with training items. The total badness of fit for any experiment was therefore computed as the weighted sum of the fit to transfer and training items:

$$B = \sum_{t \in \text{training}} \tau B_t + \sum_{t \in \text{transfer}} (1 - \tau) B_t. \qquad \text{(C2)}$$

In practice, we set $\tau = .5$ for all fits reported in this article. Badness of fit was minimized using an iterative hill-climbing algorithm, with multiple attempts made from a variety of different initial parameter values.

Received January 4, 2002
Revision received November 12, 2003
Accepted November 12, 2003 ■