Mem Cogn (2013) 41:886–896
DOI 10.3758/s13421-013-0306-9

Learning and extrapolating a periodic function

Michael L. Kalish

Published online: 9 March 2013

Abstract How people learn continuous functional relationships remains a poorly understood capacity. In this article, I argue that the mere presence of nonmonotonic extrapolation of periodic functions neither threatens existing theories of function learning nor distinguishes between them. However, I show that merely learning periodic functions is extremely difficult. It is only when stimuli are presented numerically, rather than as numberless quantities, that participants learn anything like a periodic function. In addition, I show that even then, people do not regularly extrapolate periodically. The lesson is that careful methodologies will be required to understand a psychological capacity that is as idiosyncratic as the learning of complex functions appears to be.

Keywords Function learning · Individual differences

Functions, like categories, are an essential aspect of our environment and our relationship to it. We must know how much force to use to lift a bucket as a function of how much water it holds, how long to water the lawn as a function of the day's temperature, and how slowly to drink our wine to enable our safe drive home given how long we intend to stay at the party and how much we have been eating. We all must be able to quickly learn how hard to press the gas pedal to get a given amount of acceleration each time we get into a rental car, while plumbers may learn over time to estimate how far from a junction a leak is by the sound of the turbulent water.

Theories of function learning must explain how people become sensitive to the structure of a causal relationship between metric dimensions, such as frequency and distance or temperature and water loss, from a few examples. The functional form of a causal relationship does not have to be explicitly computed by the learner; indeed, the vast majority of people who ever have and ever will learn functions cannot do even simple algebra. Theories of function learning recognize this and look to describe the cognitive mechanism that enables function learning as being one or another sort of formal system. There are generally assumed to be two possibilities for the kind of computation such a system must perform: Either it hypothesizes one of a small number of explicit functions and tunes their parameters to fit the data (e.g., Carroll, 1963; Koh & Meyer, 1991; McDaniel & Busemeyer, 2005), or it generalizes from the training examples on the basis of their similarity to novel items (Busemeyer, Byun, DeLosh, & McDaniel, 1997; DeLosh, Busemeyer, & McDaniel, 1997). Precisely which functions should be available to such a mechanism defines various parametric models, and precisely how a mechanism generalizes defines instance-based, nonparametric models. Importantly, these same two possibilities are reflected in normative approaches to the problem of regression. Parametric (e.g., Bayesian regression), on the one hand, and kernel-based (e.g., Gaussian process prediction) techniques, on the other, can both be used to approximate training data to arbitrary precision.

Parametric regression is familiar to all. Given an expression such as $f(x) = ax^2 + bx + c$, it is a simple matter to find the values of the parameters $\{a, b, c\}$ that minimize the error, $[f(x) - y]^2$ associated with the expression when given a set of (x,y) pairs. Bayesian regression (Williams, 1998) similarly uses parametric functions but predicts the posterior distribution of y given the history of (x,y) pairs by integrating over a set of candidate functions weighted by their posteriors, which are themselves dependent on both the data and the prior distribution over candidate functions. Early approaches to the cognitive processes of function learning (Carroll, 1963; Koh & Meyer, 1991) were concerned largely with identifying the set of candidate functions people (or, ambiguously, their learning mechanisms) may have access to and the facility they (or their mechanisms) have with finding the best-fitting parameter values.

Kernel-based methods of predicting y given the history of (x,y) pairs might seem generally less familiar, but psychology is replete with theories of this type. Exemplar models of

M. L. Kalish (✉)
Institute of Cognitive Science, University of Louisiana at Lafayette, Lafayette, LA 70504-3772, USA
e-mail: kalish@louisiana.edu

category learning (the GCM; Nosofsky, 1986) are kernel-based estimators of the probability density of y (a category label) given the current (metric) stimulus and the history of (x,y) pairs (Ashby & Alfonso-Reese, 1995). The kernels for the GCM are *radial basis functions*, and so novel items tend to share a category label with old items that are close by in x (we might call this a kind of *simple* similarity). That is, for Gaussian process prediction, a particular kernel that determines how responses are to be related given a particular relationship between stimuli precisely defines the concept of *similarity*. When predicting metric values, rather than category membership, kernels other than the radial basis function are often useful.[1]

The point of raising the connection between human and machine learning is that we can, in this instance, take advantage of well-established results in machine learning theory to clarify the relationship between psychological models. It has been made clear that the parametric and kernel approaches are formally identical; for any given parametric regression model, there is a perfectly matching kernel, and vice versa (see Griffiths, Lucas, Williams, & Kalish, 2009; Lucas, Griffiths, Williams, & Kalish, 2013). That is, Bayesian linear regression and prediction with Gaussian processes are just two views of the same solution to regression problems. This means that the general question of whether human learning can only be understood as using functions or similarity is essentially moot.

However, this formal isomorphism does not equate to psychological equivalence. In general, the nature of the functions or kernels required to approximate human learning is likely to differ sharply; simple similarity is not mimicked by any simple parametric function, and vice versa, except in some limited cases. Behavioral research methods are, in principle, able to distinguish under what conditions particular kernels or functions are psychologically plausible. For example, one view put forward in computational models of function learning is that when generalization occurs, it is simple and when parametric functions are employed, they are linear (Busemeyer et al., 1997; DeLosh et al., 1997; Kalish, Lewandowsky, & Kruschke, 2004). Another view is that parametric functions are always employed and their form varies from linear to exponential or logarithmic (Caroll, 1963; Koh & Meyer, 1991) or even to periodic, trigonometric, functions (Bott & Heit, 2004).[2]

Busemeyer et al. (1997) rightly emphasized the importance of extrapolation in distinguishing the underlying representation acquired in function learning. In turn, Bott and Heit (2004) pointed out that neither simple kernels nor linear functions could provide the basis for one particular form of extrapolation. Bott and Heit trained participants on a set of points carefully drawn from a periodic (cosine) function and collected transfer responses for an additional set of points (see Fig. 1). The experiment consisted of eight phases, each of which alternated a training block with a transfer block. The data from the final transfer phase were analyzed by fitting either a linear or a cosine function to each participant's responses. Bott and Heit found that 17 of 26 participants were best fit by the cosine function, and 15 of these showed a nonomonotonic pattern over the range of the transfer items. Figure 1 shows the results for these 17 participants.[3] Bott and Heit took these results as evidence against pure similarity-based associative learning processes and claimed that the results falsify even hybrid theories of function learning, such as EXAM (Busemeyer et al., 1997) and POLE (Kalish et al., 2004), both of which use similarity to govern parametric response selection.

As was discussed above, it is a mathematical truth that every possible extrapolation pattern is necessarily just as consistent with some arbitrary kernel as with some arbitrary function. The critical question is whether performance is more plausibly described by reference to a kernel or a function. To accept Bott and Heit's (2004) conclusion that participants have learned the parameters of a sinusoidal function of the stimulus value, one must observe both that the function required to account for a participant's performance is plausible and that the kernel required to do so is not. If the results shown in Fig. 1 convince the critic that participants have actually learned the same parametric sinusoidal functions that have been used by the experimenter to account for their data, then we must presume one of three things. First, we might propose that people are able to explicitly estimate the parameters of cyclic functions [i.e.,

[1] There is a close relationship between kernel-based prediction and the semivariance-based analysis presented in the Method section below. In kriging, interpolation between points is done using the empirical covariance matrix from which the semivariance is computed. Kernel-based prediction essentially uses a parametric model of the covariance between points to approximate the empirical covariance matrix; thus, it assumes a *Gaussian process*. This parameterized model is the kernel that defines the similarity between novel, extrapolation, stimuli, and items in the history of training.

[2] The distinction between these views is more nuanced than this dichotomy suggests, of course. McDaniel and Busemeyer (2005) suggested that Busemeyer et al.'s (1997) model "may capture . . . nonmonotonic prediction behavior . . . if we assume that once participants have learned the periodic nature of a mapping, they recode the inputs at the end of each period to repeat the cycle" (p. 38), and Kalish et al. (2004) observed both that "people might . . . transform [a] complex cyclic function into a simpler function with only a single cycle" (p. 1093) and that "in some contexts, cyclical functions may be simple" (p. 1094).

[3] These results are derived from a personal communication (L. Bott, January 6, 2008), since the published (Bott & Heit, 2004) table of parameters is unfortunately not formatted properly. Variability in the best-fitting parameters was not available.

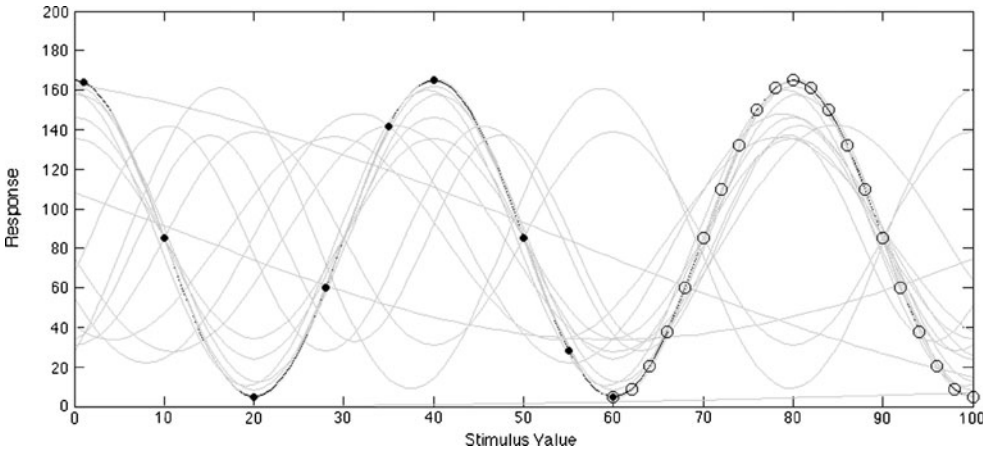

Fig. 1 Bott and Heit's (2004) results for participants classified as using a cosine function. Training stimuli are shown with solid markers, and transfer items are open markers. The dark line is the to-be-learned function, and the light gray lines are the best-fitting functions, estimated by Bott and Heit from the final transfer block for each of the individual participants classified as not using a linear function

say to themselves something like "I will respond 0.5 + 0.5 sin ($4\pi(x + .25)$)"]. Second, we might think that this parameter estimation is done by a cognitive mechanism, so that someone with no training in trigonometry might nonetheless learn a sinusoidal function. Third, we might propose that people make their responses on the basis of perceived similarity of the transfer items to the training items but that their judgments of similarity are from kernels consistent with such functions.

My suggestion is that these results are not convincing evidence of any of these possibilities. The critical feature of Fig. 1 is that the predictions of the best-fitting cosine functions used to describe the transfer responses are in direct contradiction to the to-be-learned function across most of the training range. Yet participants did quite well on the training items and had been doing well on them for at least two training phases. This inconsistency is problematic for all three of the possible accounts given above. The first hypothesis proposes that people have learned a parametric function; to account for the results, this view must hold that each person has simply forgotten what the relevant parameters are and substituted different ones for the transfer phase. Their demonstrated excellence during two successive training phases makes this kind of forgetting (of four numbers at most) rather hard to accept. The second hypothesis proposes that people's learning mechanism has done the forgetting. The utility of a mechanism that could so rapidly forget something it learned so well over so long a period is questionable at best. The third hypothesis proposes a kind of complex similarity learning, different for each participant, which would result in both the training and extrapolation responses. While this is possible, a simpler alternative might be more plausible.

Given the details of the experiment, with just one cycle of the periodic function being used at test, it is quite plausible that the curves in Fig. 1 are simply artifacts of the procedures used by Bott and Heit (2004). One cause of the results might be the choice of training and transfer items. This prevents participants from linearly extrapolating their learned responses, because linear extrapolation would require them to produce negative responses and these are disallowed. The items also influence similarity-based responses in that one kind of simple similarity would put participants in the odd situation of always choosing the response "0" since this is the criterion for the final training item. Also, participants might use overall similarity, which might dictate that they should respond "0.5" to all test items, since this is the average over all training items. Just one alternation between these two similarity-based strategies would produce the appearance of nonmonotonic responding after fitting sinusoidal models to the data. Another cause of the results in Fig. 1 might simply be the analysis used by Bott and Heit (finding the best-fitting sinusoid) combined with the method of locating all transfer items in a compact region of the stimulus space encompassing a single cycle of the criterion function. Generalizing a process from a small region to a whole space via a parametric function often produces curves that are not representative of what the process is actually doing—a lesson familiar to any who have tried to fit data using polynomial functions. These sorts of problems emphasize the difficulty in assuming that the mere presence of nonmonotonicity in the best-fitting model of a person's responses is *prima fascia* evidence for having learned a periodic function or, indeed, evidence for or against any particular theory of function learning. More work must be done to determine the nature of any particular nonmonotonicity before theories of its origin can be evaluated.

This study represents an attempt to better understand the mechanism that enables human function learning. The goal at the outset was simply to replicate the results of Bott and Heit (2004) in an experimental design with a better chance of determining a representative best-fitting function or plausible kernel when people learn a cyclic function. The narrative of the article is complicated by pilot results that suggested that getting people to learn such a function might actually be quite difficult. Thus, in addition to methodological changes designed to better characterize people's learned functions, this

article employs two factors that might make learning nonmonotonic functions less difficult: changing the nature of the stimulus presentation and increasing the prior familiarity of the criterion function.

Preliminary difficulties and design considerations

For reasons outside the scope of this article, a student in my laboratory set out to replicate the Bott and Heit (2004) design and to add some extensions (Doyle, unpublished). Bott and Heit presented their stimuli as filled horizontal bars and used two sorts of cover stories for their instructions, one of which was designed to be neutral and the other biasing toward a periodic function (they found no substantive difference between instruction conditions). Doyle replicated this design and tested an additional stimulus condition where the stimulus and the response were both vertical, rather than horizontal, and an additional biasing cover story. The results were disappointing; only 1 participant unambiguously learned the criterion values for all training stimuli.

The reasons for this failure to replicate were not clear, but two possibilities immediately present themselves. First, while Bott and Heit (2004) mentioned that they followed the procedure of DeLosh et al. (1997), their method of stimulus presentation was not entirely clear from their article. Delosh et al. employed graph-like displays in which each stimulus was presented as a labeled bar graph, with numerical tick marks identifying the approximate stimulus value. Bott and Heit's stimuli may have lacked these, as their Fig. 1 shows, but the relevant figure is not a screen shot and differs in layout from that described in their Method section. Our failed replication attempt did not have tick marks and value labels, so I manipulated this factor of stimulus presentation in the present experiment. The presence of numbers allows not only a more precise perception of the stimulus, but also a wide variety of representations (Griffiths & Kalish, 2002), reflecting a wide range of possible response strategies (Tenenbaum & Griffiths, 2001).

Second, while the difference in display details is speculative, it is incontrovertible that the participants tested in my lab and those tested by Bott and Heit (2004) differed considerably. Their participants were drawn from a population of students with substantial mathematical training and facility; the entry requirements for our respective universities are wildly different. Bott and Heit's theory holds that people employed a sine function to learn the training items; most if not all of our participants could not generate or identify a sine function. To the extent that people's strategies come into play during learning, this difference is problematic. The present experiment included a manipulation (the presence or absence of a "hint") designed to make the notion of a periodic function more available to some participants and, thus, to attempt to influence the kind of strategic processes they might use.

Questions of representation and experimental design

The experiment presented here is aimed at identifying the response profiles that arise when people learn training items drawn from a periodic function. It therefore differs from Bott and Heit's (2004) design in the placement of training and transfer items to increase statistical leverage, as well as to allow linear extrapolation, which is impossible in Bott and Heit's design. The experiment also includes a factorial combination of the presence/absence of tick marks and of a dynamic hint prior to training and, finally, displays the stimuli and responses in the manner of Kalish et al. (2004), which prevents simple segmentation of either display.

Experiment

Method

Participants

One hundred eighty-two undergraduate students from the University of Louisiana at Lafayette volunteered to participate and received partial course credit.

Stimuli

For training, 11 (x,y) pairs were selected, where x ranged from .25 to .75 in steps of .05, and the to-be-learned function was $y = 0.5 + 0.5\sin(4\pi(x + 0.25))$. The criterion values formed one full cycle of a sine curve, with the beginning and end points at the middle y value, as shown in Fig. 2. Test stimuli were the 23 values of x from [0,1] in increments of 0.0455. Of the test stimuli, only the central, $x = 0.50$, stimulus was used during training.

The physical stimulus was presented as a colored bar, oriented horizontally, whose extent signified the stimulus value. The bar varied from 0 to 34 cm in 0.34-cm increments, so stimuli were rounded to the nearest hundredth of a logical unit before conversion to the physical stimulus display. The criterion, when available, was provided as a 0- to 20-cm-high vertically oriented bar, displaced to the right of the stimulus.

Design

Participants were assigned sequentially to one of four groups that crossed the presence or absence of tick marks and value labels with the presence of absence of a

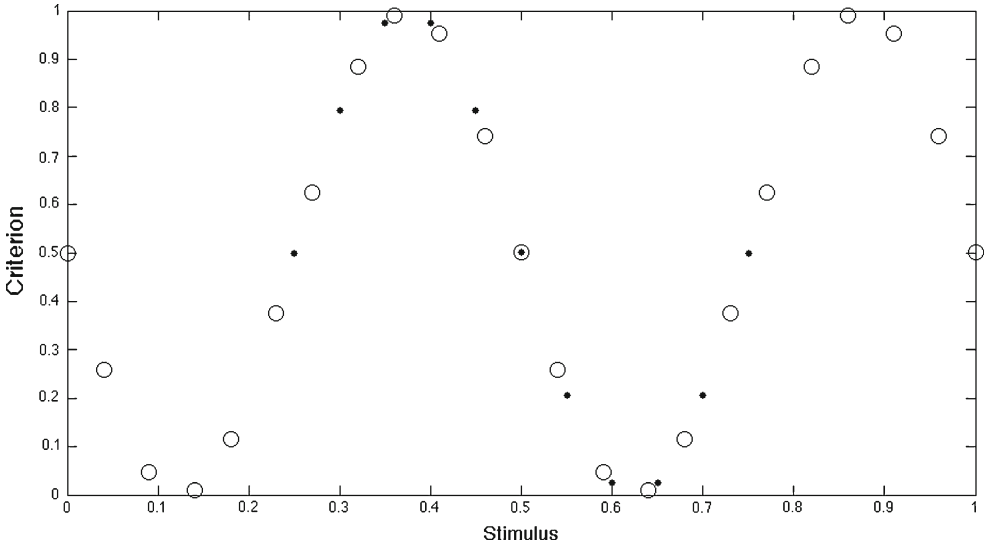

Fig. 2 The training (*filled markers*) and transfer (*open markers*) items used in the present experiment

pretraining hint (described below). Sixty-six participants were in the control (no hint, no ticks) group, 37 in the hint-only group, 35 in the ticks-only group, and 44 in the hint+ticks group.

Procedure

Participants in the hint groups read a brief description of periodic functions, using examples such as the vertical position of a spot on a bicycle wheel as it rolls horizontally. In addition, the hint included a dynamic interaction with the stimulus display; a slider was placed above the stimulus bar, and participants were allowed to move the bar back and forth from $x = 0.25$ to 0.75. As the stimulus moved, the criterion bar was continuously updated in accord with the criterion function. Participants were required to manipulate the slider through its range at least once and were allowed to do so ad lib before the first training phase. They were instructed that learning the relationship would shorten the time taken to complete the experiment.

Each participant completed two training phases, each composed of four blocks of all 11 training stimuli in random order and each followed by a test phase composed of one presentation of each of the 23 test stimuli, also in random order.

In the training and test phases, each trial began with the presentation of a stimulus bar. The participant used a vertically oriented slider, displaced to the right of the maximum stimulus extent, to produce a response. As the slider was manipulated, the participant's response was presented as a bar whose top was aligned with the current position of the slider. The participant then selected an on-screen button to indicate when he or she believed the response was correct. For training trails, feedback was then provided to the right of the response bar, via a bar whose height represented the correct response. This display lasted for 1 s unless the participant's response deviated by more than 0.03 (or ±6 mm) from the correct response, in which case he or she was required to use the slider to adjust the response until it was correct (within 2 mm of the criterion). There was a 2-s intertrial interval during which the stimulus, response, and criterion were all absent from the screen. For transfer, there was no feedback, and the intertrial interval followed selection of the response button.

Results

The results of the experiment were analyzed in three steps. First, I examined the training data to see whether participants were able to make accurate estimates of the criterion during the experiment. Then, with the results of the training sessions in mind, I analyzed the transfer phases to see how participants generalized—whether they learned the function—and, most important, I looked for evidence of periodic or nonmonotonic extrapolation.

Training

A rough estimate of learning can be found in the errors participants made. Each group's average absolute error was computed for each of the eight training blocks. I performed a Bayesian analysis[4] on the learning rates over blocks. Each participant's errors were modeled as an exponentially decreasing function of block, with an intercept (k)

[4] An excellent introduction to Bayesian analysis for psychologists is Kruschke (2011). In essence, the Bayesian approach requires formulation of a parameterized model of the data-generating process and a method of determining the posterior distribution of those parameters, given both a believable prior over possible parameter values and the data obtained from an experiment. In some cases, the posterior can be obtained algebraically, but in other cases, the data-generating process is complicated enough to require a numerical approximation of the posterior. This is obtained by using Markov chain Monte Carlo sampling.

and a decay parameter (d), and normally distributed error having precision τ : $error = ke^{-d\,\mathrm{block}} + N\left(0, \frac{1}{\sqrt{\tau}}\right)$. The parameters for each participant s in each group i were modeled as draws from gamma distributions (parameterized as shape, rate) $k_{si} \sim \Gamma(a_{ki}, b_k)$ and $d_{si} \sim \Gamma(a_{di}, b_d)$. Priors on these hyperparameters and on τ were $\Gamma(0.001, 0.001)$. The goal of the analysis is to determine the posterior distributions of intercept and decay parameters for each participant. To do so, I used a numerical method of MCMC sampling, in which the stationary distribution of a Markov chain stands in for the posterior distribution. The chains must be long enough so that any effects of initial conditions are removed (that is, there must be a *burn-in* period) and so that the distributions of samples across multiple chains are effectively unchanging (to ensure that it has actually converged). $\widehat{\mathrm{R}}$ measures the convergence of a set of MCMC samples to its stationary distribution. (Gelman & Rubin, 1992); when the measure is near 1.0, the samples are close to the posterior distribution. I used JAGS (Plummer, 2003) to collect three chains of 2,000 samples, of which the first 500 were discarded as burn-in. The $\widehat{\mathrm{R}}$ statistic for each parameter was near unity, indicating that more sampling would be unlikely to produce any change in the posterior distributions.

The mean posterior of the intercepts, which represent the error rate *before* the first trial, $\overline{x}_{ki} = \frac{a_{ki}}{b_k}$, are shown in Table 1. The hint-only group has a credibly[5] lower intercept than any other group, all of which are not credibly different. The hint gave participants a head start on learning, but only in the absence of tick marks.

The mean learning rates are $\overline{x}_{di} = \frac{a_{di}}{b_d}$; for the four groups. The learning rates for the two tick groups are credibly greater than those for the no-tick groups. The rate of the hint+tick group is also greater than that of the tick-only group, although this difference is credibly less than the difference due to ticks per se. The ticks improved learning, and the hint made this improvement greater.

An important caveat to these results is that they measure learning accuracy only in terms of overall error, not in terms of the functional form of whatever participants learned. Visual inspection of participants' last block of training responses showed what appeared to be a number of different response functions during the last block of training. For example, some participants seemed to be acting in accord with a linear function (of varying slope and intercept); other response functions appeared sigmoidal (with varying gain and location), while still others appeared possibly to be cyclic (with considerable variability in frequency and regularity of oscillation). I therefore attempted to cluster participants into groups with similar response profiles to examine both the nature of the individual variability and the role of experimental factors in it. Unfortunately, the most obvious solutions, to cluster on the basis of responses to the last block of training items or on the basis of the parameters of regression equations for each participant, produced no reliable or interpretable results. This is possibly because the inter- and intraparticipant variability makes identification of "similar" participants difficult; participants with quasi-periodic response functions that deviate only slightly from each other on a qualitative view are actually very dissimilar in quantitative terms. In an attempt to address this problem, I converted the last-block responses into semivariances and clustered on these.

The semivariance of a set of observations is a construct used frequently in geostatistics to construct and model a semivariogram that provides a graphical measure of the observations' covariance structure (Goovaerts, 1997). The semivariogram plots the change in response measure for a pair of points, the squared difference in normalized deviation $d_{ij} = 0.5\left(\frac{y_i - \overline{y}}{s} - \frac{y_j - \overline{y}}{s}\right)^2$, as a function of the absolute difference, or lag, in stimulus value $\left|x_i - x_j\right|$. Any given function will have a semivariogram of a characteristic shape. This shape is a graphical representation of the kernel that would be required for a kernel-based regression procedure to predict the response function.

For the data here, the 11 training items are evenly spaced. There are thus 10 observations of items that differ by one lag (0.34 cm of actual stimulus size), 9 that differ by two lags, and so on, down to just 1 observation at the extreme difference of 17 cm. This redundancy allows a mean semivariance to be calculated for each lag. To increase the reliability of these means, the lags were binned into six approximately equal quantiles before averaging. The resulting individual variograms were used as feature vectors for EM-clustering, which revealed three reliable clusters. Figure 3 shows that the clusters included one where participants clearly learned the periodic form of the criterion values (cluster 3), one where participants responded approximately linearly (cluster 2), and one where participants learned some but not all of the criterion's nonmontonicity (cluster 1). The ability of the EM-clustering method to find clusters and the cohesive nature of these clusters are evidence that the semivariance approach yielded sensible descriptions of participants' responses.

An analysis of the distribution of clusters by experimental group revealed that almost all the cluster 3 learners were in the two tick mark groups, as shown in Table 1. The four conditions differed in the distribution of participants in clusters, $X^2(6) = 59.99$, which gave a Bayes factor far in

[5] A *credible* value of a parameter is one that lies within the interval where 95 % of the highest density of the posterior distribution of that parameter has been estimated to be. A value outside that "HDI" is deemed *not credible* (see Kruschke, 2011). For example, a *credible difference* is one where the HDI on difference scores does not include the value zero. Credibility is similar grammatically to significance, although the two differ conceptually.

Table 1 Results for analyses of learning and transfer for each of the four conditions

	Condition			
	Control	Hint	Ticks	Hint+Ticks
Learning intercept (±95 % HDI[a])	3.901(0.194)	3.448 (0.232)	4.080 (0.303)	3.962 (0.291)
Learning rate (±95 % HDI)	0.027 (0.013)	0.018 (0.015)	0.102 (0.031)	0.142 (0.033)
Number in training clusters[b]	46, 19, 1	19, 14, 4	10, 12, 13	10, 8, 26
(percentages)	(70, 29, 1)	(51, 38, 11)	(29, 34, 37)	(23, 18, 59)
Interpolation errors (±95 % HDI)	0.30 (0.04)	0.28 (0.06)	0.21 (0.06)	0.15 (0.05)
Extrapolation errors (±95 % HDI)	0.52 (0.04)	0.52 (0.06)	0.43 (0.06)	0.40 (0.05)

[a] HDI is high density interval. The variances of the posteriors are used to compute the HDIs, since the distributions are approximately symmetric

[b] The response profiles of the extracted clusters 1–3 are shown in Fig. 3

excess of 100, making this nonuniformity definitive. The primary effect was in the difference between the conditions that did and did not have tick marks, $X^2(2) = 51.01$, again a definitive difference. The hint group differed only slightly from control; the $X^2(2) = 6.09$ marks a negligible Bayes factor of 1.28 in favor of nonuniformity. The two tick conditions did not differ substantially; the $X^2(2) = 4.16$ gave a Bayes factor of 2.29 in favor of the null, which is strong evidence that the distribution of clusters in these two conditions was uniform.

These results provide converging evidence that the ticks were a more efficient aid to learning the nonmonotonic training items than was the hint.

Transfer

Transfer items spanned a larger stimulus range than did the test items and so, potentially, provide a more sensitive measure of the extent to which participants learned a cyclic function. In addition, because transfer included both interpolation items placed within the training range and extrapolation above and below that range, the transfer data can speak to any difference between interpolation and extrapolation. That said, there are limits on what the data can reveal given the small number of observations per participant that are possible within the constraints of the experiment.

The primary question is whether or not participants learned the criterion function and applied it to extrapolation items. Visual inspection revealed that not even 1 of the 182 participants did so unambiguously. There are many possible tests for this; I chose to examine the amount of *error* (the absolute divergence from the criterion function) on extrapolation and interpolation items. A participant who learned the criterion function should treat these items the same, so I performed a hierarchical Bayesian analysis of these error rates. There are 11interpolation items (including the training item $x = 0.5$) and 12 extrapolation items. Error rates, y_i, for each item were modeled as $y_i \sim N\left(\bar{x}_s + d_s, \frac{1}{\tau_y}\right)$, where $\bar{x}_s$ is the

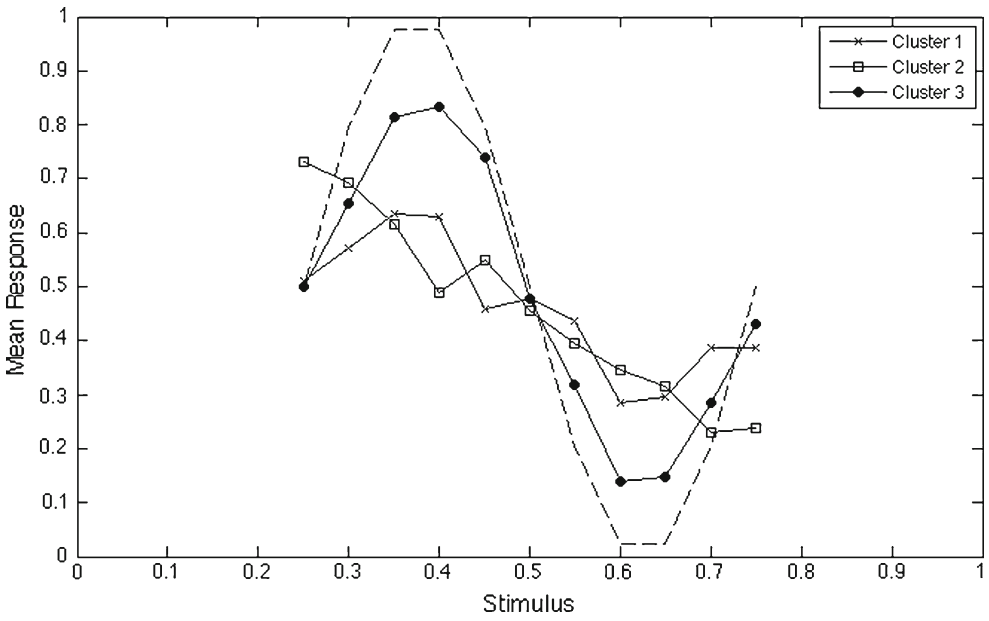

Fig. 3 The means of the responses for the participants included in each of the three clusters (the criterion function is shown as a dashed line). See Table 1 for the number of participants grouped in each cluster

mean error rate across all items for participant s, d_S is the difference (or *deflection*) in error between interpolation and extrapolation items, and τ_y is the precision of the response distribution over all participants. The participant parameters are drawn separately for each condition: $\bar{x}_s \sim N\left(\mu_{\bar{x}c}, \frac{1}{\tau_{\bar{x}}}\right)$ and $d_s \sim N\left(\mu_{dc}, \frac{1}{\tau_d}\right)$. The hyperparameters other than the mean deflections μ_{dc} were constrained to be positive and were distributed as $\Gamma(0.001,0.001)$ while the μ_{dc}s were each distributed as $N(0,0.1)$. Four chains of 1,500 samples were drawn following 1,000 burn-in steps; convergence was good, with $\widehat{R}$ measures for each parameter all near unity.

The critical result is this: No participant, either individually or as a group, could be reasonably held to be accurate on the extrapolation items (the group results are shown in Table 1). On interpolation items, the two tick groups were more accurate than the no-tick groups, and the hint+tick group was more accurate than the tick-only group. But every group was credibly less accurate on the extrapolation than on the interpolation, and no individual's error rate on extrapolation items could be said to credibly be zero. Thus, not even 1 participant credibly extrapolated the criterion function.

Discussion

This discussion has three parts. First, I summarize the main results. Second, I discuss the difference between my results and those of Bott and Heit (2004). Third, I raise the question of what participants learned if they did not learn the criterion function and whether their extrapolation could nonetheless be rational.

The data from the learning phase show that that the presence of tick marks and value labels was much more effective than the presentation of an extended, interactive hint in improving learning. While the hint provided some initial increase in accuracy in the absence of tick marks, it did little to alter the functional form people learned during training. Similarly, the hint alone also did not improve accuracy on interpolation items at transfer. The failure of training and interpolation showed clearly that, without tick marks, participants were largely unable to learn the correct responses for a nonmonotonic function. In contrast, the mere presence of tick marks and value labels improved the accuracy of both learning and interpolation. Finally, the combination of hint and ticks allowed participants to achieve the highest levels of accuracy. This is fully consistent with the notion that the tick marks and value labels provided increased precision of representation, allowing greater accuracy, and an increased representational flexibility, allowing the hint to be more effective in shaping the results of learning.

Critically, however, the variable and idiosyncratic nature of transfer, especially of extrapolation, provided little support for the idea that participants employed periodic parametric functions. Even those participants who learned the training items (cluster 3) failed to extrapolate accurately. In this regard, the results of the present experiment are fundamentally at odds with those described by Bott and Heit (2004). They reported that 17 out of 26 participants learned to use a periodic function. Not even 1 participant of the 182 in the present study showed the facility to extrapolate a periodic function cyclically. While there are differences in the stimuli used at training and test and in the number of training and test phases in the two studies, preliminary work suggested that these seem unlikely to account for the difference in results. Of the 71 participants tested in four different attempts to replicate Bott and Heit's reported method (Doyle, unpublished), only 1 produced accurate responses.

It is possible that the apparent difference in results is due, instead, to differences in analysis paired with the difference in design. Bott and Heit (2004) used the maximum likelihood parameters of a sinusoidal function to classify participants'

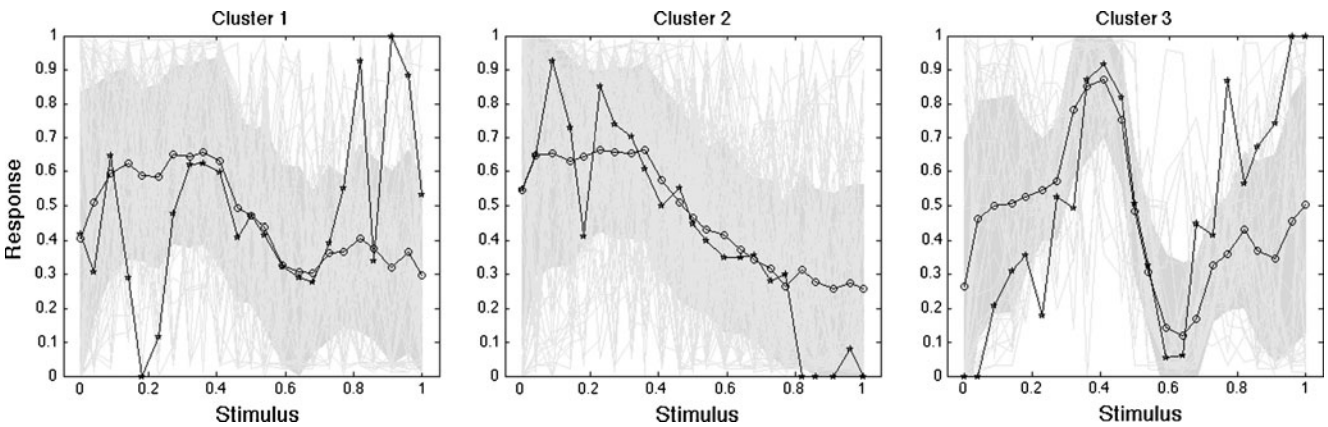

Fig. 4 Test results for each of the three clusters. The data (*faint gray lines*) are shown along with the mean responses (*circles*), their 95 % confidence intervals (*gray regions*), and the predicted means based on the estimated covariances of each cluster (*stars*)

responses as monotonic or not. However, their design provided only a limited range of data for fitting, potentially making these parameters unreliable indicators of what participants actually learned. As Fig. 1 shows, the models reported by Bott and Heit are at least partly at odds with their claim that people both learned the correct responses for the training items and generalized using a periodic function. Figure 1 shows that many of the models that fit people's extrapolation performance predict that their performance on training items should be quite different from the criterion function. Since people reportedly did learn the training items, this strongly suggests that many of these people did not abstract a periodic function, even though they may have extrapolated nonmonotonically. Indeed, when I applied this method to the Doyle (unpublished) results I found that 6 participants were classified as nonmonotonic even though only 1 of them had accurately learned the training data.

It is worth noting that applying this hierarchical regression approach to the present experiment produces just this sort of contradictory description of the results. I fit a sinusoidal and a linear regression to the two extrapolation regions ($x < 0.25$ and $x > 0.75$) of each participant and tabulated the number of such regions better fit by the sinusoid. Out of 185 participants, 105 were better fit by the sinusoid in at least one of these regions, with the sinusoid fitting better than the linear model in 68 upper regions and 67 lower regions. Only 30 participants, however, were better described by a sinusoid in both extrapolation regions. Critically, within this small set of participants, the frequencies of the best-fitting sinusoids were almost perfectly uncorrelated between the two regions, $r(30) = -.065$, 95 % high density interval (HDI) = (−0.246, 0.120) (using the method of Schisterman, Moysich, England, & Rao, 2003). The best-fitting sinusoids from each region also deviated systematically from that estimated from the interpolation data. The latter was close to the true sinusoid, with a mean frequency of 12.36 (vs. 12.57 for the criterion). The correlations of the lower and upper regions with the interpolation region were $r(67) = -.01$, 95%HDI = (−0.132, 0.113) and $r(68) = .24$, 95%HDI = (−0.351, −0.122), respectively. The negative correlation between the upper extrapolation and interpolation frequencies is credible; whatever people learned, it was not a single sine function.

It is clear that participants in the present experiment did not generalize the criterion function, did not employ any sinusoid, and did not treat the two extrapolation regions identically. What is far less clear is exactly how to characterize what participants did do. While the range of possible descriptions is potentially as broad as the range of different individuals' responses, I will end with just one question here: Is extrapolation performance at the group level predictable from training performance? Each of the three clusters extracted from the training data is characterized by a mean semivariance matrix. I used that matrix to compute a predicted transfer response profile for each cluster via Gaussian process prediction (Rasmussen & Williams, 2006). To the extent that this predicted response function lies near the observed mean responses, it is possible to say that people were generally rational; that is, their extrapolation, while not consistent with the criterion, would be consistent with their own performance during training. Figure 4 shows the results of this analysis.

The central feature of Fig. 4 is the response variability of each cluster, shown as the width of the shaded region. Each participant contributes only one response to each stimulus in this analysis, so this variability is entirely between participants. The responses to some stimulus values, especially near the extremes of the stimulus range or at inflection points of the mean response function, are strongly bimodal (beta distributions fit to these conditional response distributions have parameters <<1.0). The bimodality hints at two related features of the transfer responses. First, people largely anchor their responses at either zero or one. This is consistent with other observations in function learning (Kwantes & Neal, 2006), but in the present experiment, the choice of anchor was not consistent. The inconsistency of the anchor is mirrored by the second feature, which is the tendency of participants to make extreme quasi-periodic responses. Figure 4 also shows the individual responses of each participant. The frequent reversals of these responses suggest that the vast majority of participants interpreted the criterion's periodicity either as quasi-periodicity or simply as noise. The latter is more evident in cluster 1, where participants respond with either zero or one throughout the stimulus range, while the former seems a better characterization of cluster 3, where inflections in individual response functions are less numerous and more consistent.

The semivariance-based predictions reinforce as well the distinction between extrapolation and interpolation at test. The predictions are derived from the covariance matrices computed from the training responses, and they do a good job of predicting interpolation overall. Thus, it appears that people generalized what they learned at training to similar items at transfer, and the groupings based on training semicovariance are appropriate. However, the predictions are quite bad for extrapolation, lying at the edge or outside of the confidence intervals. This indicates that people who can be clustered together on the basis of their interpolation behavior nonetheless differ widely in their extrapolation.

The results of this experiment have been analyzed in a number of nonstandard ways. This deserves some explanation. On the one hand, Bayesian parameter estimation methods were used instead of frequentist (Fisher or Neyman–Pearson) hypothesis tests. The movement toward Bayesian statistics is a strong one in psychology, and the choice to use these methods here is unrelated to

the methodological difficulties involved in identifying what people have learned in a function-learning task. That challenge was met with a different set of nonstandard techniques, where semivariance and EM-clustering replaced parametric model fitting and likelihood-based model selection. Here, the choice of modeling techniques makes all the difference, as is pointed out in the application of likelihood-based modeling to these data above.

The present experiment bridges the methodological gap between Bott and Heit (2004) and Kalish et al. (2004) with the controlled addition of tick marks and by including both interpolation and extrapolation transfer items over a wide range of stimulus values. The presence of tick marks was not enough to entice participants to induce the criterion function, but it was critical in allowing participants to learn the training data's nonmonotonicity. The ticks might enable this learning in one of two ways. First, the ticks remove uncertainty from both the stimulus and response values, making memory a much more effective tool in learning the task. Second, the value labels that accompany the tick marks introduce a numerical stimulus coding so that mathematical operations, such as factoring or taking a modulus, become possible (Griffiths & Kalish, 2002; Tenenbaum & Griffiths 2001). The increased learning performance and interpolation accuracy afforded by the ticks suggests that the increase in precision is substantial. The failure of the ticks to induce "accurate" or cyclic extrapolation suggests that the recoding was not particularly effective for the present experiment's population. My participants do not appear to have access to the kind of deliberate strategies that Bott and Heit's (2004) participants may have used.

The message of this article, then, is essentially cautionary. Theories of function learning are abstract enough to allow description of any observed behavior, including the capacity to learn periodic functions. People with limited mathematical training seem largely unable to convincingly learn such functions from a small set of examples, even when stimuli are presented numerically and in a condition where the periodic function is clearly relevant. If the aim of function-learning research is to study the mechanisms upon which such learning is dependent (rather than, say, the way people use their knowledge of trigonometry), then we must try to characterize just what people can do and what they actually do when failing to do what we try to teach them. Individual differences in function learning make analysis of group performance difficult, and the nature of the data required to identify individual strategies makes experimental design challenging. In the realm of relatively complicated functions, it is clear that we know more about what people do not do than we know about what they can do, let alone why they do it.

Author's Note Preparation of this article was facilitated by a grant from the Louisiana Board of Regents. Thanks to Charles Barousse and Laurie Robinette for assistance with data collection and to Margery Doyle for sharing her first-year project research.

References

Ashby, F. G., & Alfonse-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

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

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

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

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

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford: OUP.

Griffiths, T., & Kalish, M. (2002). A multidimensional scaling approach to mental multiplication. *Memory and Cognition, 30,* 97–106.

Griffiths, T., Lucas, C., Williams, J., & Kalish, M. (2009). Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems, 21*.

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

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

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. New York: Academic Press.

Kwantes, P. J., & Neal, A. F. (2006). Why people underestimate *y* when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1019–1030.

Lucas, C. G., Griffiths, T. L., Williams, & J. J., Kalish, M. L. (2013). Modeling human function learning. Manuscript submitted for publication.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin and Review, 12,* 24–42.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57.

Plummer, M. (2003). JAGS: A program for the analysis of Bayesian graphical models using Gibbs sampling. www-ice.iarc.fr/~martyn/software/jags

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Boston: MIT Press.

Schisterman, E. F., Moysich, K. B., England, L. J., & Rao, M. (2003). Estimation of the correlation coefficient using the Bayesian approach and its applications for epidemiologic research. *BMC Medical Research Methodology, 3,* 1–4.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *The Behavioral and Brain Sciences, 24,* 629–641.

Williams, C. K. I. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 599–621). Cambridge: MIT Press.