

August 2017

Revisiting peak shift on an artificial dimension: Effects of stimulus variability on generalization.

E. J. Livesey

University of Sydney, Sydney, Australia

I. P. L. McLaren

University of Exeter, Exeter, UK

Correspondence:

Evan Livesey
School of Psychology
University of Sydney
Griffith Taylor Bldg A19
NSW 2006
Australia
Fax: +61 2 9036 5223
Email: evan.livesey@sydney.edu.au

Please cite as:

Livesey, E.J. and McLaren, I.P.L. (in press). Revisiting peak shift on an artificial dimension: Effects of stimulus variability on generalization. *Quarterly Journal of Experimental Psychology: Special Issue on the Psychology of Associative Learning in honour of N.J. Mackintosh*.

Abstract: One of Mackintosh's many contributions to the comparative psychology of associative learning was in developing the distinction between the mental processes responsible for learning about features and learning about relations. His research on discrimination learning and generalization served to highlight differences and commonalities in learning mechanisms across species and paradigms. In one such example, Wills and Mackintosh (1998) trained both pigeons and humans to discriminate between two categories of complex patterns comprising overlapping sets of abstract visual features. They demonstrated that pigeons and humans produced similar "peak-shifted" generalization gradients when the proportion of shared features was systemically varied across a set of transfer stimuli, providing support for an elemental feature-based analysis of discrimination and generalization. Here we report a series of experiments inspired by this work, investigating the processes involved in post-discrimination generalization in human category learning. We investigate how post-discrimination generalization is affected by variability in the spatial arrangement and probability of occurrence of the visual features, and develop an associative learning model that builds on Mackintosh's theoretical approach to elemental associative learning.

**Revisiting peak shift on an artificial dimension: Effects of stimulus variability on
generalization**

Associative learning theorists often assume that the psychological mechanisms responsible for conditioning in the laboratory animal are the same, or at least fundamentally very similar, to those governing human learning. Although we often expect core components of learned behavior, such as generalization and discrimination, to operate in similar ways across species, the translation of conditioning experiments to human learning equivalents is often a fraught exercise that can yield unexpected results. At the heart of the problem is that a healthy adult human, faced with a relatively simple learning task, can make use of a wide variety of cognitive capacities to approach, solve, or circumvent any given task. As Mackintosh put it in the 22nd Bartlett Memorial lecture, “people are not so easily persuaded to behave like pigeons, rabbits, or rats” (1995, p194). Mackintosh’s comparative research on associative learning in pigeons and humans often focused on developing novel ways to tackle this problem, with the aim of revealing the commonalities in learning across species. He argued that a relatively simple form of elemental associative learning underpinned much of human and animal behavior, a hypothesis he studied in collaboration with many postgraduate students and colleagues, with a particular focus on discrimination learning and generalization. The present study evolved as a continuation of one such line of enquiry, and concerns the nature of feature-based discrimination learning in humans.

Associative learning models essentially describe feature-based learning. They assume that the perceived properties of a stimulus come to be mentally linked with representations of the events that they predict. Mackintosh (1995; 1997; 2000) argued that this capacity to learn about basic physical properties of events is something we share with other animals, whereas the ability to compare the properties of stimuli, to identify the manner in which they are related, and to mentally represent those relations symbolically and independently of the stimuli themselves, was something uniquely human. This opinion has been shared by many researchers over the last century, but has at times been a contentious one (e.g. see Penn, Holyoak & Povinelli, 2008 and the ensuing commentaries). For reasons that will become clear, Mackintosh's work on this issue focused on discrimination learning and its impact on generalization. In particular, he pursued an interest in the phenomenon known as the peak shift effect in animals and in people.

Feature-based generalization and the peak shift effect.

As a guiding principle, feature-based learning theories predict that generalization will be monotonically related to stimulus similarity. Situations or stimuli that share many features in common with prior experience are relatively likely to elicit the same behavioural decisions, with the strongest generalization reserved for occasions where precisely the same stimuli are re-experienced, as described by Shepard's (1987) "universal law" of generalization. It is the *exceptions* to this law that are often of greatest theoretical interest and the peak shift effect is an example that is widely observed in animal discrimination learning (Ghirlanda & Enquist, 2003). When an animal is trained

to discriminate between two similar stimuli, for instance by reinforcing responses to one (S+) and not the other (S-), under some circumstances the subject no longer displays the strongest conditioned response to S+, but rather to a slightly different stimulus, one which is less similar to S-. Most animal demonstrations of this phenomenon use stimuli that lie along simple natural continua. For instance, in Hanson's original (1957; 1959) demonstration, pigeons trained on a spectral (colour) discrimination between 550 nm (green) S+ and 560 nm (yellow-green) S- showed stronger conditioned responding to a 540 nm (blue-green) test stimulus than to the well-learned S+. In most demonstrations of the peak shift effect, a discernible peak in learned behavior is clearly evident, such that responding declines at more extreme test values (e.g. 530 nm, 520, etc.), which suggests that behavior is still controlled by the presence of simple physical characteristics of the test environment.

When similar experiments are run in humans, the results can be quite different. Two-choice categorization tasks (for instance where S+ and S- are replaced with S_{Left} and S_{Right}) tend to produce gradients that are best described as sigmoidal and monotonic, instead suggestive of a simple relational response rule (see Livesey & McLaren, 2009). In contrast, Blough (1973) found that pigeons still produced peak-shifted gradients when given an equivalent two-response task. Some human studies using absolute identification (e.g. participants must respond only to stimuli identical to S+) have reported peak-shifted gradients after discrimination learning between an S+ and S-. However, these procedures tend to be highly susceptible to relative stimulus effects, for instance showing a strong sensitivity to the range of stimuli used on test. In some experiments that measure

generalization along simple continua such as stimulus luminance, test range effects completely overshadow any experience of the specific featural properties shown during training (see Thomas, 1993 for a review), which suggests that relational comparison between stimuli dominates under these conditions. A simple explanation is that participants' understanding of the dimensional construct along which the stimuli are related (e.g. colour or brightness as a continuum) comes to control the way they make decisions.¹

This raises the question; do feature-based learning mechanisms actually produce anything like the peak-shift effect in humans? Answering this question requires a different approach, one in which the dimension along which the stimuli are organized is not easily recognized. Mackintosh (1995; 1997) advocated using stimuli whose relatedness is not easy to articulate, either because it is masked by competing task goals, or because it is simply not of a form that is easy to describe. One solution used by Wills and Mackintosh (1998) was to construct complex patterns of abstract shapes, or *icons* (A. J. Wills & McLaren, 1997; Jones, Wills & McLaren, 1998; see Figure 1) and by varying the icons that appeared in each stimulus, create a series of stimuli with a systematic and ordinal

¹ As one reviewer pointed out, there are other interpretations of Thomas' range effects, especially those conducted using a luminance dimension. For instance, intensity dimensions such as luminance also produce biased gradients in animal learning (Ghirlanda & Equist, 2003) and, as Ghirlanda (2002) has shown, this may be due to the way changes in stimulus intensity are represented, and not because of an understanding of stimulus relations. Luminance perception is also likely to be particularly susceptible to sensory adaptation effects. Coupled with continued learning during extensive testing (e.g. see Ghirlanda 2007 and Livesey & McLaren, 2009), this might explain some of the strong range effects that Thomas reported. Nonetheless, the presence of range effects is a cause for concern when using these types of procedures and has contributed to the motivation to look for other solutions to studying discrimination learning and generalization.

relationship to one another, in effect an artificial continuum along which stimuli could be ordered.

[Figure 1 about here]

Wills and Mackintosh (1998) used these icon stimuli to provide a means of experimentally modeling how any given set of related stimuli might be represented in learning. They made the assumption that related stimuli activate overlapping sets of elements, an assumption that Blough (1975) had used to great effect in combination with a summed prediction error learning algorithm to quantitatively model discrimination and generalization along natural continua. The frequency of occurrence (i.e. the number of copies presented) of a particular icon within a stimulus was taken to be analogous to the level of activation of a particular set of stimulus elements. Thus it was assumed that varying the frequency of presentation of an icon in this way would have a corresponding effect on the activation of units involved in the representation of that icon. Blough's model assumed that when stimulus elements were ordered appropriately according to their dimensional properties, the pattern of activation generated by perception of a stimulus drawn from a naturally occurring dimension would approximate a normal density function or Gaussian curve. The complex stimuli used in Wills and Mackintosh's experiments were organised in exactly this way.

[Table 1 about here]

The design in Table 1 is taken from the unpublished PhD thesis of Oakeshott (2002) and is similar to other published examples. Each stimulus has four copies of a central icon, three copies of the icons immediately adjacent to the central icon, and one copy each of the next icon on either side (henceforth this will be referred to as a 1-3-4-3-1 distribution). Stimuli with the same sampling properties (i.e. the same relative proportions of adjacent icons) can be constructed for each of the stimulus positions along the artificial dimension. Thus the icon stimuli can be assumed to have overlapping activation of elements in the manner hypothesized to underpin representation of stimuli that differ along real physical dimensions. Examining Table 1, after training on the shaded S+ and S- stimuli, peak shift would be evident if the response rate to N+ was higher than the response rate to both S+ and F+.

Unlike discrimination learning with stimuli that possess simple relations, Wills and Mackintosh (1998) found that discrimination learning and subsequent generalization with these artificially constructed stimulus dimensions readily produces peak shift effects in both pigeons and in people. Indeed, several studies (McLaren & Mackintosh, 2002; Oakeshott, 2002; Wills and Mackintosh, 1998) have shown that a pseudo-Gaussian icon distribution such as the one shown below in Table 1 can be used to successfully mimic the results obtained with other animals using naturally occurring dimensions. These results have been used to lend support for the idea that feature-based associative learning operates in a similar way in humans and other animals, that post-discrimination generalization possesses the same lawful properties under at least some conditions (those in which associative learning is not in competition with relational rules). Wills and

Mackintosh (1998) further claimed that the peak-shift result was uniquely supportive of the elemental learning approach of the kind suggested by Blough (1975) among many others. For instance, they argued that Pearce's (1987; 1994) configural learning model could not easily account for the result because it predicted that generalization decrement across the dimension would be too rapid, even after extensive discrimination learning. We will return to this argument in the discussion.

Notwithstanding the demonstrations of peak shift discussed above, there appear to be versions of the icon-based paradigm that do not produce peak shift. In particular, post-discrimination gradients are affected by the spatial variability of the stimuli – the manner in which the icons are organized or distributed within the array when presented to subjects. In the icon experiments that have produced peak shift effects, the arrangement of icons has been randomly varied from one exemplar to the next. That is, on every occasion in which S+ is presented, the position of each icon within the spatial array is determined randomly. Thus each exemplar of S+ will vary in terms of the spatial organization of its icons. Oakeshott (2002) observed a difference in post-discrimination gradients produced by stimulus arrays in which the arrangement of icons was either variable as just described, or fixed so that each icon is presented in one given position. In the variable condition, there were no constraints on the position within the array at which any given icon could appear and the arrangement was randomized on each trial. In the fixed condition, the arrangement was randomized at the beginning of the experiment but then remained constant for all presentations of the stimuli. The spatial arrangements of neighbouring stimuli on the dimension were also interdependent such that the icons that

were shared with adjacent stimuli remained in the same positions across the dimension. For instance, the 8 icons that S+ and S- shared in common retained their spatial locations between stimuli. Likewise, N+ had four icon positions that differed from S+ but was otherwise identical. F+ shared only 5 icon positions in common with N+ and only 2 in common with S+.

Oakeshott (2002) found that pigeons produced strikingly different patterns of generalization when the spatial arrangement of the icons was fixed versus variable. While variable conditions generally produce peak-shifted generalization gradients, there was much stronger generalization decrement under fixed conditions, such that responding was clearly at its highest for S+. We also observed a similar pattern in a pilot study using human participants performing a categorization task (see Livesey, Pearson & McLaren, 2005; Livesey & McLaren, 2011). The absence of peak shift in the fixed condition conflicts with a simple elemental analysis that disregards the position of the icons, which would predict equivalent effects in both conditions. The difference between conditions suggests that a strong change in the form of the generalization gradient was caused by a relatively simple change to the regularity of the physical features that comprise the discriminative stimuli. That associative learning theories fail to capture this effect is symptomatic of what we see as a broader challenge to associative learning theory as it exists today. Models of associative learning rarely specify the mechanics of stimulus representation in anything but the most abstract terms, which means that many simple manipulations of basic stimulus properties, even those that seem to have a profound effect on learning-related phenomena, are simply beyond the scope of most learning

models. Theories that *have* explicitly specified how variations in simple stimulus features should manifest in stimulus representation often prove to be surprisingly effective in accounting for differences in animal learning phenomena (e.g., McLaren, Kaye and Mackintosh, 1989; Ghirlanda & Enquist, 1998).

The present study examined similar manipulations of stimulus variability in human categorization, with a view to better understanding the implications of these effects for models of feature-based learning. In order to examine the form of the generalization gradient in more detail, we used larger stimuli with richer icon-based distributions. We constructed stimuli with distributions that we have previously used to test for changes in feature associability in human discrimination learning (Livesey & McLaren, 2007). Using manipulations of the trial-to-trial variability in both the spatial layout and the frequency of occurrence of the component features, we tested for variations in generalization. After reporting the results of these tests, we will then discuss the implications of these results for the nature of stimulus representation in associative learning.

Experiments

Since the procedures for all three experiments were very similar, we describe them together, before going through the results of each. To provide some continuity with past studies, Experiment 1 replicated the peak-shifted generalization gradient previously reported when trial-to-trial spatial and frequency variability is used during training (e.g. Wills and Mackintosh, 1998; McLaren and Mackintosh, 2002), but with this larger

stimulus dimension from which we can gain a well-sampled generalization gradient.

Experiments 2 and 3 then went on to explore manipulations of spatial and frequency variability and their impact on post-discrimination generalization.

In each experiment, participants were presented with a two-choice learning task, in which they categorized each stimulus using left and right responses. Thus, rather than using an S+ and S-, we will refer to the training stimuli as S_L and S_R for stimuli requiring left and right key presses respectively. Each stimulus consisted of 36 icons arranged in a 6x6 array (see Figure 1). The dimension was designed so that the two training stimuli shared the same proportion of features in common as did the 12-icon training stimuli used by Oakeshott (2002). However the size of the array allowed stimuli centred on adjacent icons to be much more similar (in terms of the proportion of features that they shared in common), thereby providing a greater number of measurable steps over which to gauge post-discrimination gradients. Hence two steps along the dimension, rather than one, separated S_L and S_R . Table 2 shows the 15 stimuli along the full dimension. The design of this full dimension is symmetrical around stimulus 8, with stimuli 7 & 9 used as the training stimuli S_L and S_R respectively. For all analyses, the two ‘sides’ of the dimension were collapsed, with results to be expressed as test accuracy plotted over the ordinal distance to the nearest training stimulus. This has been the convention used in several previous studies involving human categorization (e.g. Jones and McLaren, 1999; Livesey & McLaren, 2009; McLaren and Mackintosh, 2002; Natal, McLaren, & Livesey, 2013; Wills and Mackintosh, 1998), and simplifies analysis of the gradient over multiple points.

Combined in this way, S_L and S_R will be referred to as S , and thus values 1 and 15 are 6 steps from S , values 2 and 14 are 5 steps from S , and so on.

[Table 2 about here]

The icons were randomly ordered along an artificial dimension that determined which icons appeared in each stimulus array. There were 24 types of icons used to represent the elemental components of the dimension, labeled by a letter A-X in Table 2. Each stimulus consisted of a variable number of copies of 10 types of icon, with the icons positioned consecutively along the artificial dimension. The stimulus was composed in such a way that the number of copies of each icon reflected its ordinal position on a Gaussian curve. This approximation to a Gaussian distribution was designed to be similar to the distribution used by Oakeshott (2002), though the stimulus itself is much larger. Each stimulus consisted of 6 of each of 2 ‘central’ icons, then progressively less copies of adjacent icons as one moves away from the centre of the distribution. The identity of each icon depended upon the position of the stimulus along the dimension. For instance, stimulus 7 (S_L) would contain 6 copies of icons K and L, 5 copies of J and M, 4 of I and N, 2 of H and O, and 1 copy of G and P. Any pair of stimuli with this 1-2-4-5-6-6-5-4-2-1 distribution that are *two* steps removed from each other along the dimension share two thirds of their icons, as do any pair of *adjacent* stimuli on the 1-3-4-3-1 distribution used by Oakeshott (2002).

Participants were first given discrimination training in which they were required to respond differentially to intermixed presentations of S_L and S_R (7 and 9 on the full dimension in Table 2). A transfer test then followed in which all 15 stimuli along the dimension were presented.

In Experiment 1 the configuration of icons used for a given stimulus could vary stochastically in terms of both position of the icons and frequency of their occurrence. This was done by sampling from a pool of icons assigned to a given stimulus such that the probability that an icon was selected depended on its frequency in the distribution for that stimulus. So, taking our example of stimulus #7, the probability that icon K would be selected was 6/36 on every sample. Sampling continued until 36 icons had been selected to make up that stimulus, and these were then randomly allocated to positions in the array. Hence both the exact frequency and the spatial position of the icons were stochastic, whilst still reflecting the distribution of icons for a stimulus at that position on the dimension. Experiments 2 and 3 used manipulations of the spatial arrangement of the icons within exemplars of each stimulus in order to test for a similar effect of spatial variability as that reported by Oakeshott (2002). For participants in the spatially variable condition, the position of each icon was randomized on every trial as already described, whereas for participants in the spatially fixed conditions, the positions of icons within the array remained constant throughout the experiment (though the initial positions chosen were random).

In Experiment 2, all participants underwent training with stimuli that contained no trial-to-trial frequency variability, that is the exact number of copies of each icon as indicated in Table 2 were used every time a stimulus was presented, even if the location of each icon was randomized. In Experiment 3, spatial variability and frequency variability were manipulated independently. The method used to create stochastic stimuli in Experiment 1 (i.e. selection with replacement) had to be modified to accommodate fixed spatial conditions. Thus, on every trial in the frequency variable conditions, a proportion of the icons for a given training stimulus were replaced with icons from the corresponding locations in the other training stimuli. Thus, in the spatially fixed, frequency variable condition, there was variation from trial to trial in the exact number of each icon, but the spatial structure of the patterns remained as constant as possible, and no icon appeared "out of place" in the array.

We will first outline the methods and results from each experiment, including a curve-fitting analysis of the generalization gradient in each experimental condition. We will then discuss the theoretical implications of the research for associative learning theories.

Methods

Participants and Apparatus

In Experiment 1, 16 undergraduate students participated in the experiment, and were given monetary compensation for their participation. The experiment was run using a Macintosh desktop computer attached to a 17-inch CRT monitor. Participants were tested individually in a dimly lit room. In Experiment 2, a total of 99 students

participated, 43 tested individually in the same fashion as Experiment 1, and 56 as part of two practical classes. Group testing in the practical classes was conducted using Macintosh iMac computers with 15-inch color monitors, in a large classroom with computers spaced approximately 1m apart. In Experiment 3, 65 students participated in the experiment under the same individual testing conditions as used for Experiment 1.

We applied a standard learning criterion (e.g. Livesey & McLaren, 2007) to exclude participants who did not show evidence of learning during training. Participants who did not score higher than 55% accuracy on the icon stimuli over the second half of training were removed from the analysis. Four participants in Experiment 1, 35 participants (11 individually tested and 24 group tested) in Experiment 2, and 17 participants in Experiment 3 did not pass this criterion and were removed from the analysis. This left N=12 in Experiment 1, N=64 in Experiment 2; 33 in Group Fixed (17 group tested, 16 individually tested), and 31 in Group Variable (15 group tested, 16 individually tested), and N=48 in Experiment 3 (12 in each of four groups).

Presentation of stimuli and measurement of responses was programmed using REALbasic software in all experiments. Participants made left and right key responses by pressing 'x' or '.' respectively on a standard computer keyboard.

Stimuli

Each icon stimulus consisted of an array of 36 icons presented against a black background. On screen, each icon appeared within an area approximately 8 mm wide by 8 mm high, the array of 6x6 icons measuring approximately 5 cm x 5 cm. The stimulus

array was surrounded by a thin square white border. As described above, the icons with which each stimulus was endowed depended on the position of the stimulus on the artificial dimension (Table 2). 24 icons from a pool of 36 were assigned to icon positions A to X so that each participant would in fact be presented with a different set of randomly ordered icons.

Presentations of the icon stimuli were interleaved with filler trials consisting of uniform colored squares that differed in hue. The colors were varied in similar fashion to the artificial dimension, with two very similar shades of green presented during training, S_L and S_R , and a wider range of colors, varying from blue-green to yellow-green, presented during the transfer test. The primary purpose of the color trials was to negate any effect of immediate contrast that could influence learning of the icon discrimination.

Spatial variability: All three experiments contained at least one condition in which the spatial layout of the icons varied randomly from trial to trial during training. Experiments 2 and 3 also contained conditions in which there was no spatial variability during training. In these spatially fixed conditions, the position of each icon was randomised on the first trial for each subject, but then remained constant for all subsequent presentations of that stimulus. In these conditions, the icons that two neighbouring stimuli (i.e. stimulus n and stimulus $n+1$) shared in common appeared in identical locations in both stimuli. For instance if S_L contained 4 copies of a particular icon and S_R contained 6 copies (i.e. icon N from Table 2), then the 4 locations occupied by copies of that icon in S_L would also contain copies of the same icon in S_R . The additional two copies of the

icon in S_R would be placed in positions vacated by other icons no longer present in S_R . In contrast, in the spatially variable conditions there was no such relationship since the position of the icons varied randomly on every presentation. Examples of fixed and variable spatial conditions are shown in Figure 1.

Frequency variability: Experiments 1 and 3 used conditions in which discrimination learning contained trial-to-trial variability in the numbers of icons presented. In Experiment 1, icon presentation during discrimination training varied from trial to trial in a stochastic fashion. For instance, for stimulus S_L , a pool of 36 icons was constructed according to its distribution in Table 1 (i.e. 1 copy of icon G, 2 of H, 4 of I, etc.). To create an exemplar of S_L , each position within the stimulus array was randomly allocated an icon from the pool, irrespective of the icons selected at other points in the array. In other words, the icons were sampled from the pool *with* replacement. Take icon K from Table 2, for instance, which ideally occurs on 6 of 36 locations ($1/6$) within S_L and 4 of 36 locations ($1/9$) within S_R . Under stochastic presentation, for each location in a given presentation of S_L there would be $1/6$ probability of presenting icon K, regardless of which icons were shown in the other positions in the array.

In Experiment 3, where the presence of spatial and frequency variability were manipulated independently, stochastic presentation of the icons was achieved in a different way. In the conditions that contained frequency variability, on each training trial, six randomly selected icons in the training stimulus were swapped with the six icons from the same positions in the opponent training stimulus. Thus, the icons within each

stimulus in these conditions varied stochastically, independently of whether they varied spatially.

In Experiment 2 and in the frequency fixed conditions in Experiment 3, presentation of icons during training was not stochastic, with the precise numbers of icons always adhering to those shown in Table 2. In other words, the icons were sampled from the stimulus pool *without* replacement until they were all used. For example, there were always 6 presentations of icon K in each presentation S_L and 4 presentations of icon K in each presentation of S_R .

Procedure

On starting the experiment, participants were given written instructions informing them they would be presented with various stimuli, one at a time, and that their task was to learn which of two responses was appropriate for each stimulus. A brief description of the icon-based and hue-based stimuli was given, along with instructions about the responses, the feedback given after each response, and the time limit on each trial. Participants were told that the correct response depended entirely on the visual appearance of each stimulus, and to attend to the whole stimulus as attending to only part of it would make the later phase of the experiment more difficult to solve. They were informed that a test phase with no response feedback would be given at the end of the experiment. Participants were also informed that the icon and colored square stimuli were completely unrelated and to treat them as independent tasks.

Discrimination training: Training consisted of alternating presentations of icon stimuli and colored square stimuli. Trial order was randomized within blocks of 12 trials, containing three presentations each of S_L and S_R and three each of the corresponding filler trials (which were also reliably associated with Left and Right responses), with the condition that trials alternate between icon stimuli and filler stimuli. There were 12 blocks of training in Experiments 1 and 3, and 8 blocks in Experiment 2. In each experiment, these trials were presented as two continuous blocks of trials with participants instructed to take a short rest between the two blocks.

On each trial, participants were given 4s to respond to each stimulus. Feedback after each response consisted of a 'correct' or 'wrong' message appearing in the centre of the screen, followed by an inter-trial interval of 1.5s. Failure to respond in 4s resulted in the trial timing out and a 'no response' message appearing as feedback.

Transfer test: The 15 icon stimuli shown in Table 2 were tested, with an equal number of colored square filler trials, once again with alternating icon and filler trials. Trial order was randomised within blocks of 30 trials (one trial of each stimulus and 15 filler stimuli). In Experiment 1, there were 3 blocks (90 trials) in total whereas in Experiments 2 and 3, there were 6 blocks (180 trials). In all experiments, the test stimuli did not contain frequency variability. That is, icons that comprised the transfer stimuli were sampled *without* replacement such that the number of icons for each stimulus adhered precisely to the frequencies in Table 2. The spatial variability present in training was retained at test; spatially variable for Experiment 1 and the spatially variable conditions

in Experiments 2 & 3, and spatially fixed for participants in the spatially fixed conditions in Experiments 2 & 3.

Data Analyses.

As noted, we followed a convention established by Wills & Mackintosh (1998) to simplify analysis of the generalization gradient. To calculate response accuracy on test, a left response was considered correct for stimulus positions 1-7 and a right response was considered correct for positions 9-15 (stimulus 8 was omitted from analyses). We then collapsed the test dimension on itself and arranged 7 stimulus values classified according to ordinal distance from the nearest trained stimulus. Hence, the trained stimuli 7 and 9 form a single stimulus value 0, stimuli 6 and 10 form a single stimulus value 1 and so on up to the extreme stimuli 1 and 15 (stimulus value 6).

The curvature of the resulting generalization gradient is the critical measure of interest and may take several forms. In each experiment the 7 stimulus values were subjected to a repeated measures ANOVA to confirm that test accuracy varied as a function of stimulus value. In Experiment 2 and 3, this analysis also included the group factors of spatial variability and (in Experiment 3 only) frequency variability to test whether the pattern of generalization was affected by variability during training. The existing evidence suggests that we should expect a peak-shifted gradient, at least after training with variable stimulus conditions, but that this peak shift may be missing or reduced when conditions are spatially fixed. A key prediction that follows from Oakeshott's (2002) research in pigeons

is that the fixed condition will peak closer to (possibly precisely at) the trained stimuli. This prediction is not adequately tested by looking at trend analyses within the ANOVA. For instance unimodal non-monotonic gradients that are peak-shifted and monotonic gradients that are highest at either stimulus value 0 or stimulus value 6 can generate strong quadratic trends, see Livesey & McLaren, 2009. Thus, a different approach is necessary.

To test the hypothesis that generalization followed a peak-shifted function and that this effect was more pronounced under variable than under fixed conditions, we fit a normal curve to each individual's generalization results using maximum likelihood estimation. We assumed that this function followed the following equation, which describes a Gaussian function with baseline performance at chance:

Equation 1.

$$p(\text{correct}|x, \theta) = 0.5 + (c - 0.5)e^{-\frac{(x-a)^2}{2b^2}}$$

In this equation, x corresponds to the stimulus value (0-6) around which the test stimulus is constructed, c corresponds to the height of performance at its peak, a corresponds to the location of that peak, and b corresponds to the width of the curve (a and b correspond to the mean and standard deviation of a normal curve). For the purposes of constraining the curve fitting to realistic parameters, we made the following assumptions:

- 1) Peak location, a , could take only an integer value between 0 and 6, so that the peak of the function resided precisely at one of the 7 discrete positions along the artificial stimulus value continuum. Since our dimension is artificial and arbitrarily arranged, it is only meaningful to consider generalization peaking at a value that can actually be constructed on that dimension.
- 2) Width of the curve, b , can take a value between 0.1 (very sharply decreasing accuracy around the peak) to 20 (virtually flat accuracy function across the 7 test points).
- 3) Maximum accuracy, c , was always equal to the maximum observed level of accuracy for that individual at any stimulus value.

Thus for each participant, the function fitting returns values for one fixed parameter c and two free parameters a and b . Examples of these functions are shown for data from a real participant in Figure 2.

[Figure 2 about here]

Using this function fitting approach, we estimated the most likely location of the peak of the generalization gradient for each individual and for responses aggregated across the whole group. We used this information to compare the effects of manipulating frequency and location variability across the three experiments.

Summary. All participants in this study engaged in trial-and-error learning over multiple presentations of S_L and S_R (interleaved with another class of very different filler stimuli). For all participants, S_L and S_R were complex arrays composed of overlapping sets of features (icons), where the frequency of occurrence of each feature was determined by the position of these stimuli along an artificial dimension, as shown in Table 2. What differed between conditions was the variability in the positioning of these features within the array across exemplars, and variability in the precise number of each feature within the array across exemplars. We then tested participants' responses to related stimuli with different frequencies of icons, again determined by their position along the artificial dimension, and fitted unimodal functions to the resulting generalization gradient to determine the most likely position of the peak of this gradient at a participant and group level.

Results

Response Accuracy in Training.

Figure 3 shows performance over the course of training for each condition in the three experiments, averaged into blocks of equal numbers of trials (resulting in 6 blocks for Experiments 1 & 3, 4 blocks for Experiment 2). Gradual improvement across the phase is evident in all cases. In Experiment 1, there was a significant linear trend across blocks, $F(1,11) = 33.67, p < .001, \eta_p^2 = .754$. In Experiment 2, there was a significant linear trend across blocks, $F(1,62) = 121.59, p < .001, \eta_p^2 = .662$, which did not interact with group, $F < 0.1$. However, a main effect of group indicated that the spatially fixed group performed significantly better than the spatially variable, $F(1,62) = 14.20, p < .001, \eta_p^2 = .186$.

[Figure 3 about here]

In Experiment 3 also, it appears that variability was associated with lower accuracy in training. Main effects revealed that spatially fixed groups performed better than spatially varied, $F(1,44) = 7.82, p = .008, \eta_p^2 = .151$ and frequency fixed groups performed better than frequency variable, $F(1,44) = 8.73, p = .005, \eta_p^2 = .166$. There was also a significant linear trend across training block, $F(1,44) = 118.09, p < .001, \eta_p^2 = .729$ indicating clear improvement with training. No interactions between spatial and frequency variability were significant, though the interaction between linear trend across blocks and frequency variability, $F(1,44) = 3.90, p = .054, \eta_p^2 = .081$, and the three-way interaction, $F(1,44) = 3.79, p = .058, \eta_p^2 = .079$, approached conventional levels of statistical significance, suggesting that the condition with both spatial and frequency variability improved less than the other three groups. Taken together, these results suggest that adding variability during training reduced accuracy, which is to be expected from a very simple learning perspective.

Response Accuracy during Transfer Test. Figure 4 shows mean accuracy during the transfer test for each experiment, in each case collapsed across the two sides of the dimension to form 7 stimulus values. Across the three experiments, the conditions containing spatial variability show a consistent pattern where, moving progressively further away from the trained stimuli, accuracy increases slightly, peaks at stimulus value 3, then decreases. The conditions that were spatially fixed revealed a subtly different

pattern whereby accuracy tends to peak closer to the trained stimuli (numerically highest at stimulus values ranging from 0-2). There was a significant main effect of stimulus value in Experiment 1, $F(6,66) = 2.90$, $p = .014$, $\eta_p^2 = .014$,² and in Experiment 2, $F(6,372) = 15.17$, $p < .001$, $\eta_p^2 = .197$. Although the main effect of group in Experiment 2 was not significant, $F(1,62) = 1.40$, $p = .242$, $\eta_p^2 = .022$, a significant interaction was found between group and stimulus value, $F(6,372) = 2.90$, $p = .009$, $\eta_p^2 = .045$ suggesting that the pattern of generalization differed as a function of spatial variability.

[Figure 4 about here]

In Experiment 3, there was a significant main effect of stimulus value, $F(6,264) = 21.87$, $p < .001$, $\eta_p^2 = .332$, a significant interaction between stimulus value and spatial variability, $F(6,264) = 3.25$, $p = .004$, $\eta_p^2 = .069$, but no interaction between stimulus value and frequency variability $F(6,264) = .55$, $p = .768$, $\eta_p^2 = .012$. The three way interaction and main effects of spatial and frequency variability were also non-significant, $F_s < 1$. This suggests that the pattern of generalization was substantially affected by spatial variability but not frequency variability, despite both spatial and frequency variability clearly affecting training.

² Wills and Mackintosh (1998) focused on whether peak responding was reliably greater than responding for the trained stimuli and responding at the test extremes, whereas in this study, we focus more on differences in generalization as a function of variability and the relative position of the peak along the dimension. Nevertheless, it is worth noting that Experiment 1 replicated Wills & Mackintosh's peak shift effect in the sense that peak accuracy at stimulus value 3 was significantly greater than accuracy for the trained stimulus value 0, $t(11) = 3.07$, $p = .011$ and stimulus value 6, $t(11) = 3.77$, $p = .003$.

Peak of the generalization gradient. In order to estimate how the position of the peak of the generalization gradient is affected by variability, we fit the Gaussian function described in Equation 1 to the experimental data, in effect calculating a log likelihood function for each of the 124 individual participants in this study. From this likelihood function, one can estimate the most likely position of peak accuracy. However, the certainty of this estimate naturally differs from one participant to the next. A small proportion of participants had very flat generalization gradients (usually either at ceiling or at chance) and were best fit with a very broad Gaussian function. For instance, when $b=13$, the greatest deviation the function can predict (when $c=1$ and $a=0$ or $a=6$) across the possible x test values is 5%. Fitting functions that are this broad is not informative because a function centered on any value of a will do a reasonable job of capturing the (flat) gradient. There are several ways to mitigate this problem when attempting to estimate the most likely position of the peak of the generalization gradient. We report two here, one at an individual participant level, another at a group level, which treat functions with the 7 possible discrete values of a (i.e. stimulus values 0 to 6) as 7 independent models, and maximizing the log likelihood function for each of these models by varying the parameter b . We then compared the fit using a Bayesian Information Criterion (BIC), following this equation:

Equation 2.

$$BIC = -2\ln L + k\ln(2)$$

where $\ln L$ is the maximized value of the log likelihood function produced from Equation 1, k is the number of free parameters (b was the only free parameter, fit separately for each participant), and n equals the number of observations, which could be up to 42 per participant in Experiment 1 and 84 per participant in Experiments 2 & 3 (we omitted the rare trials where the participant failed to respond completely). BIC is a criterion commonly used for model selection, in which the lowest BIC indicates the best match with observed data.

Importantly a difference in BIC between models of less than 2 is generally considered uninformative in terms of which model provides a better fit. Thus if a participant has a minimum BIC for a particular value of a , there may be other values of a that produce virtually just as good matches to the observed data. Across all experiments, 50 participants had minimum BICs that were unequivocal (difference in BIC was greater than 2 for all other values of a). 29, 19, and 14 participants had comparable BICs for 1, 2, and 3 other values of a , respectively. 12 participants had comparable BICs for 4 or more other values of a (these are participants for whom the model fitting exercise is more or less completely uninformative). Figure 5 shows the number of participants for whom the most likely estimate of a is 0 to 6, removing those participants who had a b estimate of 13 or greater (very flat gradient) and those who had equivalent BICs for an additional 4 or more values of a in addition to the most likely value ($n = 13$ in total). These frequency histograms aggregate data for the spatially variable and spatially fixed conditions separately. For many participants in both conditions, the most likely location of peak accuracy is at stimulus value 1. However, the distributions around this value are

strikingly different, with $a = 0$ being the most likely peak for a larger proportion of participants under Spatially fixed conditions than Spatially variable. Collapsed across experiments (and complementing the analyses reported earlier), Chi Square tests reveal that a differs significantly as a function of spatial variability, $\chi^2(5, N = 111) = 12.18, p = .032$, but not as a function of frequency variability, $\chi^2(5, N = 111) = 3.09, p = .685$.

[Figure 5 about here]

A complementary analysis can be performed by assuming a equals a single value for each experimental condition and, for each a , calculating the likelihood functions across all individuals, allowing b to vary for each participant. The BICs for models with a equal to 0 to 6 is shown in Table 3. Across the four Spatially varied groups, the $a = 3$ models wins out three times and $a = 2$ wins for the Experiment 3 spatially varied, frequency varied condition. Across the three Spatially fixed groups, $a = 0$, $a = 1$, and $a = 2$ all win out once.

[Table 3 about here]

On balance, these results suggest that there is a strong likelihood that the gradients produced under spatial variability are peak-shifted and, although there is some variation in the precise location of peak performance, the peak tends to be further removed under spatially variable than under spatially fixed conditions. Evidence for peak shift under spatially fixed conditions is more equivocal. Many participants are best fit by functions

that assume a small shift in peak accuracy, as did the spatially fixed conditions in Experiment 3 taken as a whole.

Discussion

Experiment 1 demonstrated the classic peak shift effect using stimuli that varied in terms of their spatial arrangement and the frequency of occurrence of the component features of the stimulus. In Experiments 2 and 3, significant differences between the post-discrimination gradients produced under fixed versus variable spatial arrangement were evident, and all analyses suggest that this was due to the generalization gradient peaking closer to the trained stimuli under spatially fixed conditions. In contrast, the manipulations of frequency variability did not appear to have a substantial effect on the generalization gradients even though they impacted training accuracy.

Whereas the differences in generalization gradient produced under spatially fixed and variable conditions appear to be more of scale than kind in these experiments, Oakeshott's (2002) variable and fixed post-discrimination gradients appeared to be starkly different from one another. The variable condition resulted in a pronounced peak shift effect while the fixed condition displayed highest responding for S+ and sharp generalization decrement to N+ and beyond. However, Oakeshott's results were observed over relatively few test stimuli over which the gradient could be sampled. In contrast, our experiments that use larger stimuli with more gradual change in stimulus value across the artificial dimension suggest a more subtle quantitative influence of spatial variability on generalization. Although generalization gradients under spatially fixed conditions

were narrower, we still observed at least some evidence of a peak shift effect and the gradients across three spatially fixed conditions consistently possessed a negatively accelerated form, with fairly similar levels of performance for stimulus values 0 – 2 followed by more pronounced generalization decrement thereafter.

An elemental approach to modeling discrimination and generalization.

Mackintosh favored an explanation of peak shift in terms of elemental associative learning, in which the representation of a stimulus comprises a collection of smaller mental components that can individually enter into associations. These components may or may not have a one-to-one mapping to isolable features of the actual stimuli, but their mental activation is assumed to vary lawfully with changes in sampled physical properties. Learning of the general form typified by Mackintosh (1975) and Rescorla and Wagner (1972), among many others that incorporate some form of prediction error signal, has of course become a mainstay of contemporary associative theory. A pertinent example is the model of operant discrimination and generalization proposed by Blough (1975), which incorporates summed-error learning (Rescorla & Wagner, 1972) with a simple elemental mechanism for stimulus representation and generalization, to simulate discrimination learning and post-discrimination behaviour. Within its own ambit, Blough's (1975) model proved to be particularly powerful, and its central assumptions have strongly influenced more recent elemental associative models (e.g. Ghirlanda & Enquist, 1998; McLaren and Mackintosh, 2000, 2002). Blough's model predicts the general form of post-discrimination generalization gradients, and particularly peak shift, with impressive accuracy.

As noted, Wills and Mackintosh (1998) also argued that their results were less consistent with configural learning theories of the variety proposed by Pearce (1987; 1994), whereby a mental representation of the unique combination of stimuli present in a given instance (rather than individual components) enters into association. They argued that the Pearce model cannot account for a peak-shifted gradient, essentially because it predicted much stronger generalization decrement between neighbouring stimuli on the artificial dimension. For instance in reference to the simple animal conditioning example shown in Table 1, the model would predict that there is insufficient excitatory generalization from S+ to N+ to sustain strong responding and also insufficient inhibitory generalization from S- to S+ to reduce responding to S+ to a lower level than would be observed for N+ (see Wills & Mackintosh, 1998 for a full explanation of this prediction).

The merits of this argument rest on several assumptions about how stimulus elements are mentally represented in learning and are certainly questionable in light of the generalization gradients observed by Oakeshott (2002), Livesey et al., (2005) and here under fixed spatial conditions. However, we will refrain from providing a full configural learning analysis because the distinction between elemental and configural learning may not actually be particularly meaningful in this context. For instance, Ghirlanda (2015) has argued that for every elemental learning model (at least within the class of associative learning models under scrutiny here) there is a configural model that will generate the same predictions such that the two are formally equivalent. It thus seems likely that if an elemental model provides a strong quantitative account of the data, so too can a

configural learning model. Other quantitative characteristics of particular models, for instance the manner in which representations of features are normalised when they are experienced in combination with others, may be much more important than whether the learning itself occurs to elemental or configural representations (Thorwart, Uengoer, Livesey and Harris, 2016).

Instead, we will focus here on how an elemental associative model might capture the observed effects of spatial variability on discrimination and generalization. Most associative learning models do not make explicit assumptions about stimulus representation in a way that provides for such an explanation. Blough's (1975) assumptions about the underlying activation of elements, which emulate simple sensory tuning curves, seem plausible when modeling generalization between stimuli that possess continuous physical characteristics such as colour or luminance. However, continuing to use the same dimensional assumptions when the dimension of interest is completely artificial is less tenable. In any case, Blough's scheme does not speak to within-stimulus spatial variability in any obvious way. An alternative approach that has been used specifically for these artificially dimensional stimuli has been to assume that different units are solely activated by a single type of icon, essentially a one-to-one mapping of stimulus feature to representational element, with the activation of each element being directly proportional to the number of copies of each icon present in the stimulus. However, this approach assumes that the spatial location of each feature within the stimulus array is irrelevant and therefore cannot account for effects of spatial variability.

We have previously suggested a means of capturing spatial representation in the learning of these icon-based stimuli, in a way that might prove generally useful for problems involving considerations of both feature identity and spatial layout (Livesey & McLaren, 2011). There is clearly a case for a model that makes predictions which are sensitive to the effects of spatial variability and that can accommodate the results of both Oakeshott (2002) and the present experiments. But, in order for the model to be generally applicable, it needs to make minimal and generalisable assumptions about featural and spatial representation. The elemental model that we proposed then built upon assumptions about stimulus representation outlined by McLaren and Mackintosh (2000; 2002; see also Jones and McLaren, 1999; Ghirlanda & Enquist, 1999); namely distributed and overlapping representation of stimulus features, and nonlinearities in the activation of units that represent the stimuli coupled with an error-correction learning algorithm. Here we offer a simpler version of this model and test its ability to produce the generalization gradients from our three experiments.

Summary of model. We use an elemental prediction-error model with two important stages of processing; elemental stimulus representation (stimulus inputs) and output activation. Associative weights between the representational elements and the outputs are updated according to a prediction error algorithm, and predictions of each outcome on each trial are estimated using the relative activations of the outputs. These aspects of the model are widely used in other associative models. The key aspect of this model is the manner in which the features of the stimulus (i.e. the icons) activate the stimulus inputs. The stimulus inputs have overlapping and diverse sensitivities in terms of both the region

of the stimulus and the types of features that result in input activation. Each input unit is assumed to have a receptive field of a random size and shape, to be activated by a random subset of the icons, and to have variable sensitivity (i.e. it could take many or few of the right kinds of icon within the receptive field to yield strong activation). These details and the formal equations of the model are described further below.

[Figure 6 about here]

Stimulus Input. The stimuli present on a given trial are represented as a distributed pattern of activation over a set of input units. We impose relatively few constraints on what form these inputs might take other than that they are overlapping and diverse. There is no explicit representation of the artificial continuum built into the model. Rather, we assume partial spatial and feature specificity in the representational elements within the network. That is, each element is activated by several features occurring within a region of adjacent locations, and the specificity of this activation varies across elements. This is an important departure from the simpler approaches used in the past because it allows complex representation of spatially-specific features without combinatorial explosion, which could be a substantial problem if every possible feature in every possible location were represented independently. Following McLaren and Mackintosh (2002), element activation A is a nonlinear function of the sum of all the features present in the stimulus to which the element is sensitive, or total feature input F_i .

Equation 3.

$$A_i = \frac{e^{F_i} - 1}{e^{F_i} + D_i}$$

Equation 3 describes a sigmoidal function in which D_i controls how responsive the element is to the detection of features; when D_i is relatively high, it takes several features to activate the element, when D_i is relatively low, one feature in the receptive field is sufficient to achieve strong activation of the element. D_i is randomly chosen for each element, in a range from 0.1 up to the number of spatial locations to which the element is sensitive (e.g. if a particular representational element samples from an area covering 9

different locations than D for that element was randomly assigned a value from 0.1 to 9.0). Values are randomly assigned using a uniform distribution between these limits. This effectively means the representations produced by the model are not tightly controlled but are diverse and complex.

In our previous model (Livesey & McLaren, 2011), we also included inhibitory relationships between features and elements such that the presence of some features could deactivate a representational element that would otherwise respond to other features of the stimulus. A combination of inhibitory and excitatory links serves to produce *replaced* elements; some of the elements activated by feature X will be turned off and others will be turned on when X appears alongside feature Y. Thus the inclusion of inhibitory links serves a similar function to the explicit implementation of replaced or inhibited elements (see Wagner, 2003; Wagner & Brandon, 2001). Here, we have opted for an even simpler approach to test the capabilities of a model that simply contains elements that are activated by multiple features. If an element is activated by feature X and feature Y (within a given region of the stimulus) then input from X and Y may produce nonlinear activation. That is, activation by X and Y together may be more or less than activation by X alone plus activation by Y alone. However, the activation from these inputs will always be monotonically increasing. That is, activation by X and Y together will never be less than activation by X alone.

Learning. Simulating categorization experiments with associative learning requires at a minimum two output units corresponding to the left and right key press responses.

Associative weights linking the representational elements to these output units were initially set to zero and were modified using a simple error correction algorithm shown in Equation 4.

Equation 4.

$$\Delta W_{ij} = SA_i(E - I)$$

Here ΔW_{ij} is the change in the associative weight that links element i to output unit j and is changed as a function of the discrepancy between external input E (i.e. the actual outcome) and summed internal input I (i.e. the predicted outcome), weighted by the activation A_i of element i and a learning rate parameter S . Internal input I is calculated according to Equation 5, which is a simple summed prediction weighted by the activation of each element (Blough, 1975).

Equation 5.

$$I = \sum A_i W_{ij}$$

During training, one output unit (e.g. corresponding to Left category) was given an external input $E = 1$ for reinforced S_L trials and $E = 0$ for S_R trials, and vice versa for the other output unit. This learning rule is equivalent to that used by Blough (1975) and, although simplified, forms the basis for weight change in the McLaren and Mackintosh (2000; 2002) model.

Network Output. An exponential version of Luce's (1959) ratio rule was used to predict the probability of making the correct response for each test stimulus, according to Equation 6:

Equation 6.

$$p(\text{Correct}) = \frac{e^{kI_{\text{correct}}}}{(e^{kI_{\text{correct}}} + e^{kI_{\text{incorrect}}})}$$

Here, I_{correct} corresponds to the summed input I to the Left output unit for stimuli 1 to 7 and the Right output unit for stimuli 9 to 15 (with $I_{\text{incorrect}}$ corresponding to the alternate output unit in each case). A single k value was used for all subjects in each experiment, rather than being varied for each individual subject. While this parameter was used to fit the simulation data to the overall level of accuracy, it appeared to have very little effect on either the curvature of the gradients or the *relative* levels of accuracy of each condition.

Simulations. Figure 7 shows results from simulations with the model described above. These results were produced by taking the mean of 30 simulated participants for each condition across the three experiments. Each simulation used 100 representational elements. Each element was activated by icons in a receptive field ranging randomly from 1-5 icons wide and 1-5 icons high (i.e. receptive field varied randomly from 1 to 25 icons), located randomly around the stimulus display. On average, each element was activated by one third of the different varieties of icon, that is, $p(\text{icon type } x \text{ activates element } i) = 0.333$. This value was chosen as a compromise between featural specificity

and representational efficiency. After some rough parameter fitting, we chose a learning rate constant $S = .01$ for all runs in all experiments, but used a different ratio rule constant for each experiment ($k = 5.90, 2.87, 2.56$ for Experiments 1 - 3 respectively). The constant k changes the overall level of accuracy achieved during test but has relatively little impact on the shape of the generalization gradient itself. This helps to account for differences in accuracy between experiments that may well be attributable to time of testing or variations in the recruitment of motivated participants. By keeping the other parameters fixed across all conditions, we hope to provide a clear illustration of the model's strengths and weaknesses.

[Figure 7 about here]

Several characteristics of the model predictions evident in Figure 6 are worth noting. First, the model clearly predicts a peak shift effect, but more importantly one that is more pronounced under spatially variable than spatially fixed conditions. This fits well with the generalization gradients observed over Experiments 1-3. Second, the model generally predicts higher accuracy under spatially fixed conditions, especially for stimulus values 0-3. For these fixed conditions, the decline in accuracy for stimulus values greater than 3 is more pronounced in the experimental data than in the simulated data. Nevertheless, the model anticipates a sharper decline in accuracy at the test extrema under spatially fixed compared to variable conditions. Third, frequency variability during training has a negligible effect on the shape of the generalization gradient, as seems to be the case in the

experimental data. These properties hold true for a wide range of parameters that we have investigated.

The quantitative fit of the model is limited in some respects, for instance it slightly (but consistently) underestimates performance for the trained stimuli at stimulus value 0 relative to peak performance. The model also often predicts peak accuracy at stimulus value 2 rather than 3 under spatially variable conditions, however this may not be particularly problematic given the individual variability in the peak of individual gradients (illustrated in Figure 5).

This model is a simplified version of one that we developed earlier, which accounts for gradients produced in conditioning by pigeons (Livesey & McLaren, 2011). Both of these models have been loosely based on the elemental approach to stimulus representation outlined by McLaren and Mackintosh (2000; 2002). One of the main differences is that in the previous model the presence of certain icons inhibited the activation of elements within the representation of the stimulus, whereas this model relies solely on summative excitatory input. Although the inclusion of inhibitory inputs is certainly plausible, we chose the current model to provide a simpler test of the elemental approach to stimulus representation, and to provide an existence proof that this fairly limited set of associative resources would nevertheless prove equal to the task of capturing the basic pattern present in our data. The combination of a basic summation of stimulus inputs with a nonlinear activation function has been shown to be surprisingly powerful in learning complex discriminations (e.g. see Livesey, Thorwart, & Harris, 2011; McLaren &

Mackintosh, 2002 for simple examples). We are confident that other means of achieving similar forms of complex elemental representation would prove effective to this end as well (e.g. Ghirlanda & Enquist, 1998; 2005; Harris & Livesey, 2010; Thorwart, Livesey & Harris, 2012).

Elemental prediction-error learning is at its most powerful when the stimulus representations implemented in the model are rich and capture many diverse characteristics of the stimuli. The learning algorithm allows the most predictive statistics to incrementally gain control of discriminative choices. Provided the representation of the stimulus is rich enough to capture spatially specific and spatially general feature properties, different aspects will win out during spatially fixed versus variable training. In both cases the strongest learning will occur to the sub-set of elements that are more active for one training stimulus than for the other. For instance, the strongest associations with the left response will accrue to those elements for which the *difference* in activation for S_{left} versus S_{right} is the greatest. However, this discriminating sub-set will be different under spatially fixed versus variable conditions. In the fixed case, the activation across elements is highly consistent across trials of the same stimulus. Even if some frequency variability is present, as in Experiment 3, patterns of activation will be very similar across trials because icons consistently occur in the same locations. The elements that come to control the discrimination will be those that differentially respond to a highly location-specific set of features, for instance an element that responds to a particular feature in the top-left of the stimulus, which is present in S_{left} but not S_{right} . In the spatially variable case, the activation of elements across trials will be much less consistent. There will still

be elements that, for instance, respond more to S_{left} than to S_{right} on average. However, on any given S_{left} trial, not all of these elements will be active. Therefore discriminative control comes to be distributed over more elements, many of which will also be activated by other test stimuli. It is this characteristic that leads to a broader generalization gradient, which results in a peak shift being observed at a greater distance along the artificial dimension under spatially variable conditions.

This general quality of error correcting learning, the ability to extract the most predictive components of a stimulus representation, has been a major reason for its success and widespread application to associative learning. The McLaren and Mackintosh (2000; 2002) model explicitly appealed to this capability in developing a general approach to stimulus representation, one which uses rich and overlapping elemental representation of stimuli, even those lying on continua. Here we have incorporated a simple form of spatial specificity into this scheme, one which we hope can be applied to other associative learning phenomena that seem to be affected by spatial layout (e.g. Glautier, 2002; Livesey & Boakes, 2004). The scheme is also in keeping with the elemental analysis that Mackintosh appealed to in explaining the peak shift phenomenon (Mackintosh, 1995; 1997), but provides a means of accounting for differences due to variability during training, as found in this study, as well as solving other complex nonlinear discriminations.

So far, we have not discussed the role of attention in this task, even though discrimination learning is thought to affect (and be affected by) selective attention (Mackintosh, 1975).

For instance, in human learning, subjects appear to devote greater attention to the most informative or predictive stimulus features and less attention to those that are irrelevant, and we have previously found evidence that this true for these complex icon-based stimuli (Livesey & McLaren, 2007). In this study, participants were trained on a discrimination equivalent to the variable conditions used in Experiment 1, such that the most predictive features were those icons that appeared relatively often in S_{left} but rarely in S_{right} (or vice versa) regardless of their exact locations. In two experiments, participants found a subsequent discrimination easier if these predictive icons were once again the most predictive features, relative to conditions in which the most predictive icons became less predictive in the second discrimination.

In the case of the spatially fixed location conditions, selective attention may shift towards particularly predictive locations rather than (or as well as) towards predictive features. Other learning tasks indicate that participants will deliberately and strategically bias their attention towards regions of a stimulus array that are likely to contain task-relevant information (e.g. some perceptual learning phenomena, see Jones & Dwyer, 2012; Wang, Lavis, Hall & Mitchell, 2012). As noted in the methods, participants in these experiments were told to attend to the whole stimulus because attending to only one part would make the later phase of the experiment more difficult to solve. However, we cannot verify whether they followed this instruction. Even if they attempted to, associative learning may guide attention to informative stimulus locations incidentally and automatically. Implicit learning effects in visual search (i.e. *contextual cuing*; Chun & Jiang, 1998) demonstrate this capacity, albeit under different task requirements. In contextual cuing,

visual search is faster if the distractors predict the location of the response-relevant target, irrespective of its identity. This effect appears to be both incidental in terms of the sampling of predictive information (Beesley, Hanafi, Vadillo, Shanks, & Livesey, in press) and may even be relatively nonconscious under some circumstances (e.g. see Colagiuri & Livesey, 2016), but has a demonstrable effect on how the complex visual search array is processed.

Changes in selective attention towards predictive features and predictive locations are thus both possible in this task and may well impact on generalization to different test stimuli. Selective attention was not incorporated into our relatively simple model. Although the model fairs relatively well without it, we see this as a possible extension of the model's operations that is very much in keeping with Mackintosh's views on associative learning, and one which may generate further testable predictions.

Conclusion

Mackintosh (1995; 1997; Wills & Mackintosh, 1998) argued that using peak shift along an artificial dimension is one way of studying feature-based generalization in humans while circumventing the use of strategies based on relational learning. It is therefore important that our models of associative learning, which are designed specifically to predict feature-based generalization, can adequately account for the results from such experiments. The current study provides further confirmation that spatial variability within complex visual stimuli has a clear and robust effect on post-discrimination generalization. However, in this study, the fixed and variable conditions produced

gradients that appear to have a similar general curvature; all were negatively accelerated and were modestly peak-shifted for the majority of participants. Thus the gradient differences under fixed and variable spatial conditions appear to be more quantitative than qualitative. Training under spatially fixed conditions produced test accuracy generalization that, on average, peaked closer to the trained stimuli. In contrast, frequency variability during training seemed to have very little effect on the generalization gradient at test but did impact accuracy during training. These effects are accounted for by an associative learning model that incorporates a simple means of representing spatial and feature properties of the stimuli, using distributed elemental representation of stimulus properties.

References

- Beesley, T., Hanafi, G., Vadillo, M. A., Shanks, D. R. & Livesey, E. J. (in press). Selective attention in contextual cuing is driven by properties of the search task and not by the predictiveness of distractors. *Journal of Experimental Psychology: Learning, Memory & Cognition*.
- Blough, D. S. (1973). Two-way generalization peak shift after two-key training in the pigeon. *Animal Learning & Behavior*, 1(3), 171-174.
- Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, 1, 3-21.
- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36, 28-71.
- Colagiuri, B. C., Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin & Review*, 23, 1996-2009.
- Ghirlanda, S., & Enquist, M. (1998). Artificial neural networks as models of stimulus control. *Animal Behaviour*, 56, 1383-1389.

- Ghirlanda, S., & Enquist, M. (1999). The geometry of stimulus control. *Animal Behaviour*, 58(4), 695-706.
- Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, 66, 15-36.
- Ghirlanda, S., & Enquist, M. (2005). *Neural Networks and Animal Behavior*. Princeton University Press.
- Ghirlanda, S., & Enquist, M. (2007). How training and testing histories affect generalization: a test of simple neural networks. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1479), 449-454.
- Ghirlanda, S. (2015). On elemental and configural models of associative learning. *Journal of Mathematical Psychology*, 64, 8-16.
- Glautier, S. (2002). Spatial separation of target and competitor cues enhances blocking of human causality judgements. *Quarterly Journal of Experimental Psychology*, 55B, 121-135.
- Hanson, H. M. (1957). Discrimination training effect on stimulus generalization gradient for spectrum stimuli. *Science*, 125(3253), 888-889.
- Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, 58, 321-334.
- Harris, J. A. & Livesey, E. J. (2010). An Attention-Modulated Associative Network. *Learning & Behavior*, 38, 1-26.
- Jones, F., & McLaren, I. P. L. (1999). Rules and associations. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Jones, F., Wills, A.J., and McLaren, I.P.L. (1998). Perceptual categorisation: connectionist modelling and decision rules. *Quarterly Journal of Experimental Psychology*, 51B, 33-58.
- Jones, S. P., & Dwyer, D. M. (2013). Perceptual learning with complex visual stimuli is based on location, rather than content, of discriminating features. *Journal of Experimental Psychology: Animal Behavior Processes*, 39, 152-165.
- Livesey, E. J. & Boakes, R. A. (2004). Outcome additivity, elemental processing and blocking in human causality judgements. *Quarterly Journal of Experimental Psychology*, 57B, 361-379.
- Livesey, E. J., & McLaren, I. P. L. (2007). Elemental associability changes in human discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 33, 148-159.
- Livesey, E. J. & McLaren, I. P. L. (2009). Discrimination and Generalization Along a Simple Dimension: Peak Shift and Rule-Governed Responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 554-565.

- Livesey, E. J. & McLaren, I. P. L. (2011). An elemental model of associative learning and memory, in E. Pothos & A. J. Wills (Eds). *Formal Approaches in Categorization*. Cambridge University Press (pp. 153-172).
- Livesey, E. J., Pearson, L. S., & McLaren, I. P. L. (2005). Spatial variability and peak shift: A challenge for elemental associative learning. In *Proceedings of the XXVIIth Annual Convention of the Cognitive Science Society* (pp. 1302-1307), Hillsdale, NJ: Lawrence Erlbaum Associates.
- Livesey, E. J., Thorwart, A., & Harris, J. A. (2011). Comparing positive and negative patterning in human learning. *Quarterly Journal of Experimental Psychology*, *64*, 2316–2333.
- Luce, R. D. (1959). Individual choice behavior. New York: Wiley.
- Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.
- Mackintosh, N. J. (1995). Categorization by people and pigeons: The twenty-second Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, *48*, 193-214.
- Mackintosh, N. J. (1997). Has the wheel turned full circle? Fifty years of learning theory, 1946–1996. *The Quarterly Journal of Experimental Psychology: 50A*, 879-898.
- Mackintosh, N. J. (2000). Abstraction and Discrimination. In C. M. Heyes & L. Huber (Eds.), *The evolution of cognition* (pp. 123-142). Cambridge MA: MIT Press.
- McLaren, I.P.L., Kaye, H. and Mackintosh, N.J. (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.) *Parallel Distributed Processing - Implications for Psychology and Neurobiology*. Oxford. OUP.
- McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*, 211-246.
- McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*, 177-200.
- Natal, S. D. C., McLaren, I. P. L. & Livesey, E. J. (2013). Generalisation of Feature- and Rule-Based Learning in the Categorization of Dimensional Stimuli: Evidence for Dual Processes Under Cognitive Control. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*, 140-151.
- Oakeshott, S. M. (2002). *Peak shift: An elemental vs a configural analysis*. Unpublished PhD, University of Cambridge, Cambridge.
- Pearce, J. M. (1987). A model of stimulus generalisation for Pavlovian conditioning. *Psychological Review*, *94*, 61-73.

- Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109-130.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.
- Thomas, D. R. (1993). A model for adaptation-level effects on stimulus generalization. *Psychological Review*, *100*(4), 658.
- Thorwart, A., Livesey, E. J. & Harris, J. A. (2012). Normalisation between stimulus elements in a model of Pavlovian conditioning: Showjumping on an elemental horse. *Learning & Behavior*, *40*, 334-346.
- Thorwart, A., Uengoer, M., Livesey, E. J. & Harris, J. A. (2016). Summation effects in human learning: evidence from patterning discriminations in goal-tracking. *Quarterly Journal of Experimental Psychology*. First online: 29 Apr 2016, DOI: 10.1080/17470218.2016.1184290
- Wang, T., Lavis, Y., Hall, G., & Mitchell, C. J. (2012). Location and salience of unique features in human perceptual learning. *Journal of Experimental Psychology-Animal Behavior Processes*, *38*, 407-418.
- Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology*, *56B*, 7-29.
- Wagner, A. R., & Brandon, S. E. (2001). A componential theory of Pavlovian conditioning. In R. Mowrer, & S. Klein (Eds.), *Handbook of contemporary learning theories*, (pp. 23-63). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Wills, S., & Mackintosh, N. J. (1998). Peak shift on an artificial dimension. *Quarterly Journal of Experimental Psychology Section B- Comparative and Physiological Psychology*, *51*, 1-32.
- Wills, A.J., and McLaren, I.P.L. (1997). Generalisation in human category learning: a connectionist account of differences in gradient after discriminative and non-discriminative training. *Quarterly Journal of Experimental Psychology*, *50A*, 607-30.

Tables

Table 1. An example of stimuli designed with a pseudo-Gaussian distribution on an artificial dimension. This distribution of icons is that used by Oakeshott (2002; Experiment 4). Stimuli shown in bold were presented during discrimination training.

		Features comprising artificial dimension.								
		A	B	C	D	E	F	G	H	I
Stimulus values along artificial dimension.	F+	1	3	4	3	1				
			1	3	4	3	1			
	N+			1	3	4	3	1		
	S+				1	3	4	3	1	
	S-					1	3	4	3	1

Table 2. The distribution of icons for each stimulus value. The bolded stimuli, 7 and 9, were used in discrimination training as S_L and S_R respectively. For each participant, elements A to X were randomly allocated one of 36 abstract icons. All fifteen stimuli were presented during the transfer test.

Stimulus Value	Test Stimulus	Features comprising artificial dimension.																							
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
6	(F _L) #1	1	2	4	5	6	6	5	4	2	1														
5	#2		1	2	4	5	6	6	5	4	2	1													
4	#3			1	2	4	5	6	6	5	4	2	1												
3	#4				1	2	4	5	6	6	5	4	2	1											
2	(N _L) #5					1	2	4	5	6	6	5	4	2	1										
1	#6						1	2	4	5	6	6	5	4	2	1									
0	(S _L) #7							1	2	4	5	6	6	5	4	2	1								
	#8								1	2	4	5	6	6	5	4	2	1							
0	(S _R) #9									1	2	4	5	6	6	5	4	2	1						
1	#10										1	2	4	5	6	6	5	4	2	1					
2	(N _L) #11											1	2	4	5	6	6	5	4	2	1				
3	#12												1	2	4	5	6	6	5	4	2	1			
4	#13													1	2	4	5	6	6	5	4	2	1		
5	#14														1	2	4	5	6	6	5	4	2	1	
6	(F _L) #15															1	2	4	5	6	6	5	4	2	1

Table 3. BIC for models of the generalization gradient following Equation 1, where peak accuracy (a) is assumed to fall on exactly one stimulus value (0-6) for all participants, but the width of the generalization gradient (b) is a free parameter for each participant. Bolded BICs indicate the best model (i.e. most likely value of a) for each experimental condition. Experimental conditions are organized according to spatial variability, then frequency variability.

Experiment & Condition	k participants; n observations	BIC for each value of a							
		0	1	2	3	4	5	6	
1 Sp. V, Fr. V	k = 12, n = 502	966.8	860.5	777.0	552.9	852.8	1059.5	1218.8	
3 Sp. V, Fr. V	k = 12, n = 990	1388.3	1539.0	1192.3	1210.6	1797.3	2084.4	2938.0	
2 Sp. V, Fr. F	k = 31, n = 2562	4100.1	3998.5	3891.0	3701.8	4457.6	5109.2	5421.0	
3 Sp. V, Fr. F	k = 12, n = 969	1195.6	1130.1	1008.4	945.5	1362.6	2832.2	2959.6	
3 Sp. F, Fr. V	k = 12, n = 992	1301.5	1183.6	1160.3	1616.2	2026.5	2934.9	3426.3	
2 Sp. F, Fr. F	k = 33, n = 2744	3430.3	4495.4	3843.1	4930.0	7647.3	9303.5	9638.9	
3 Sp. F, Fr. F	k = 12, n = 990	1423.1	1319.7	1522.5	1974.6	2919.7	3765.1	4002.9	

Note: Sp. V and Sp. F denote Spatially varied and Spatially fixed respectively.

Fr. V and Fr. F denote Frequency varied and Frequency fixed respectively.

Figures

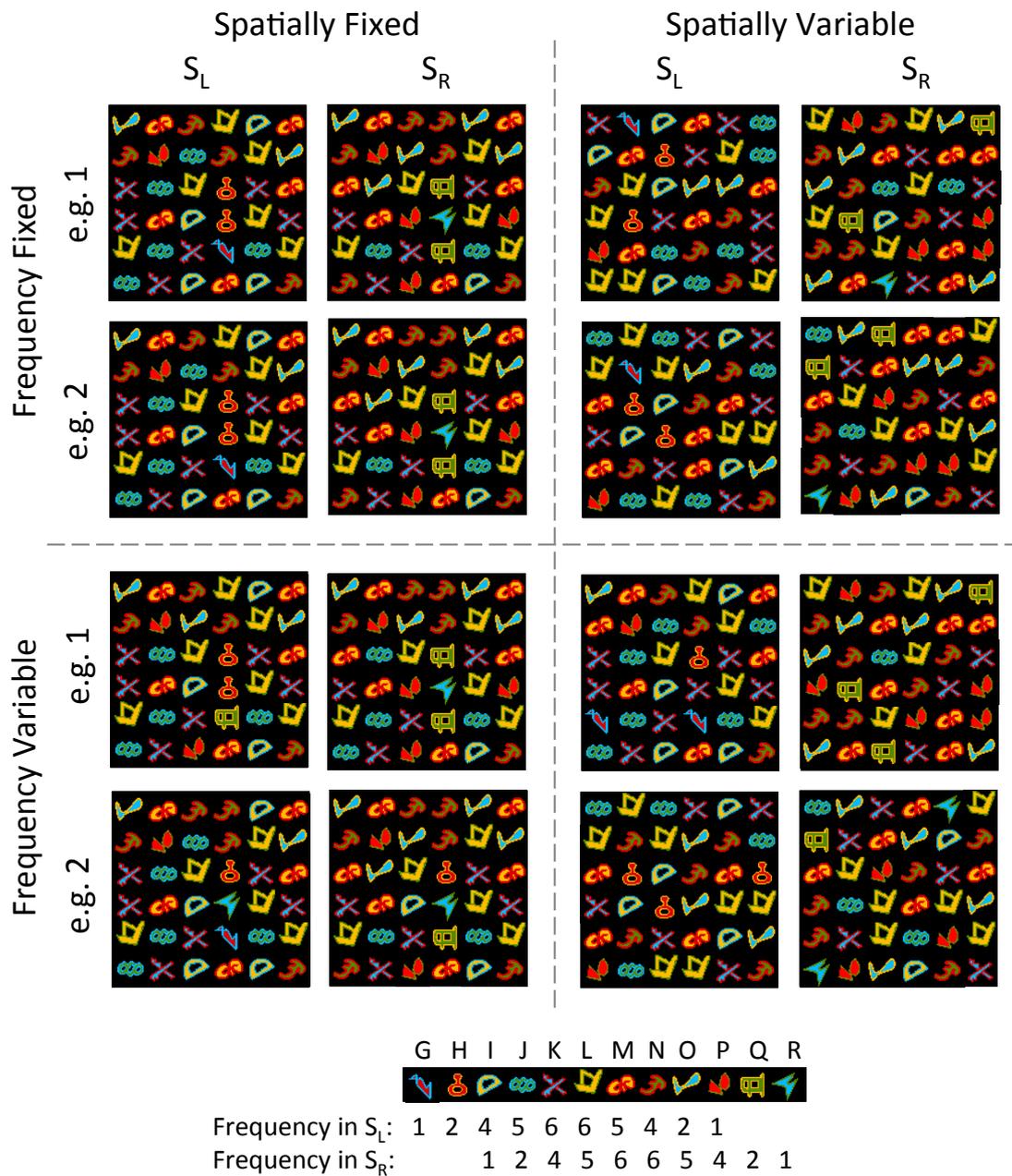


Figure 1. Examples of icon-based stimuli used to generate peak shift along an artificial dimension, given one possible ordering of the icons (i.e. the random allocation of icons G-R, shown at the bottom of the figure). Two examples of each of the S_L and S_R training stimuli are shown, for each combination of frequency and spatial variability. The icons vary slightly in color and form from those used by Wills and Mackintosh (1998) and by Oakeshott (2002).

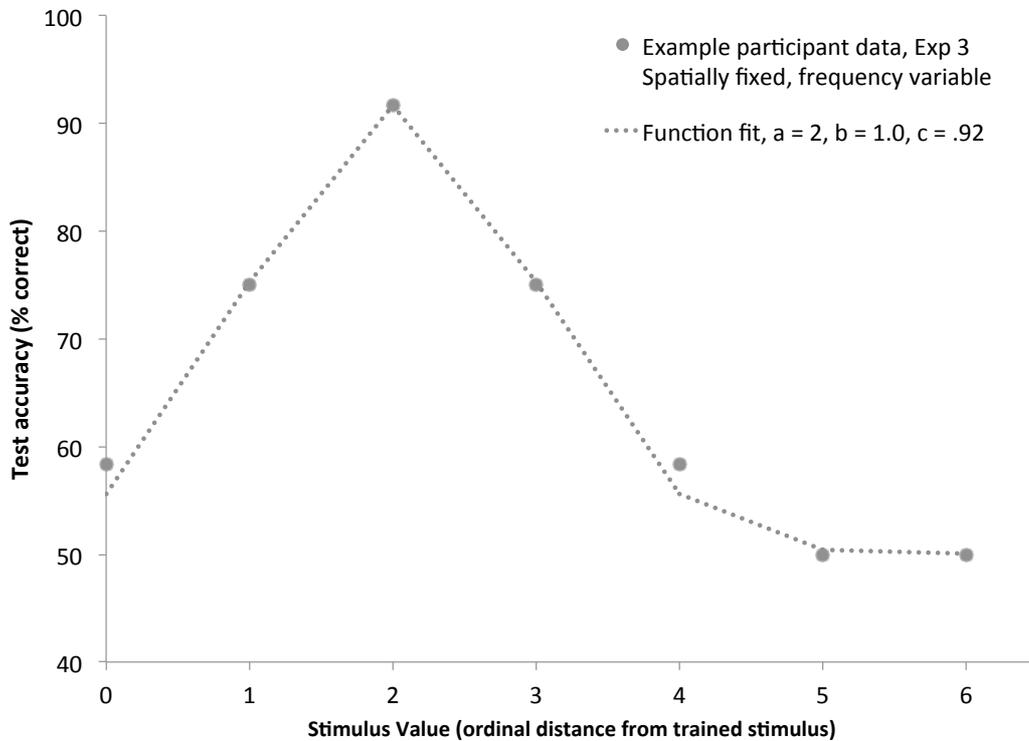


Figure 2. Example of the function fitting used to estimate the peak of gradients for individual participants. Data points represent test accuracy data for a single participant, the dotted line indicates the predicted gradient for the best-fitting function. For fitting the function, the location of the peak of the gradient a was constrained to fall precisely on one of the 7 stimulus values (0-6) and the maximum accuracy c was matched to the maximum observed accuracy. The width of the gradient b was varied continuously.

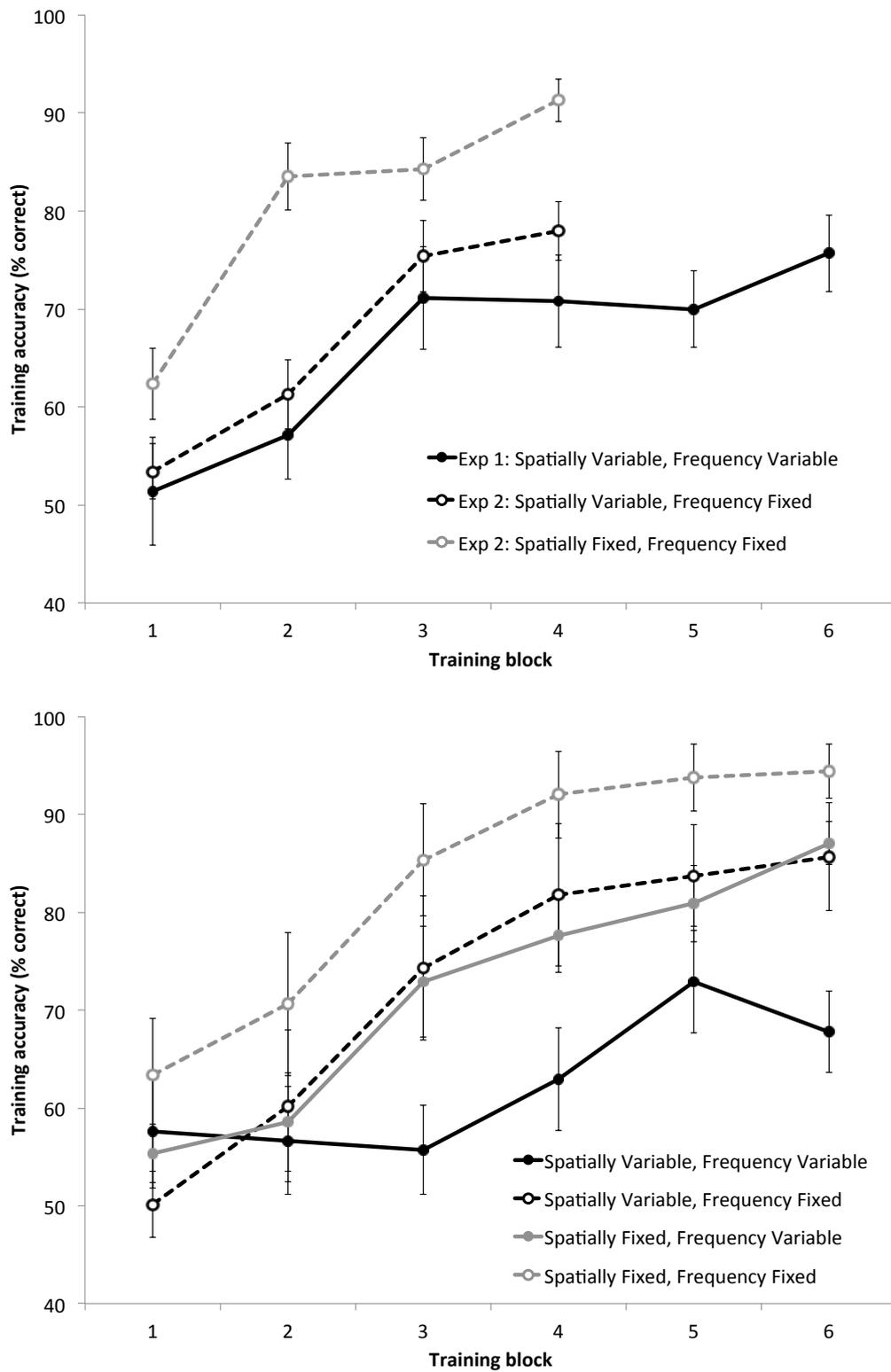


Figure 3. Percentage correct responses during discrimination learning phase. Data points indicate accuracy for successive blocks of 12 icon-based discrimination trials. Top panel shows training for Experiment 1 and the spatially fixed and variable groups in Experiment 2. Bottom panel shows training accuracy for Experiment 3. Error bars indicate SEM.

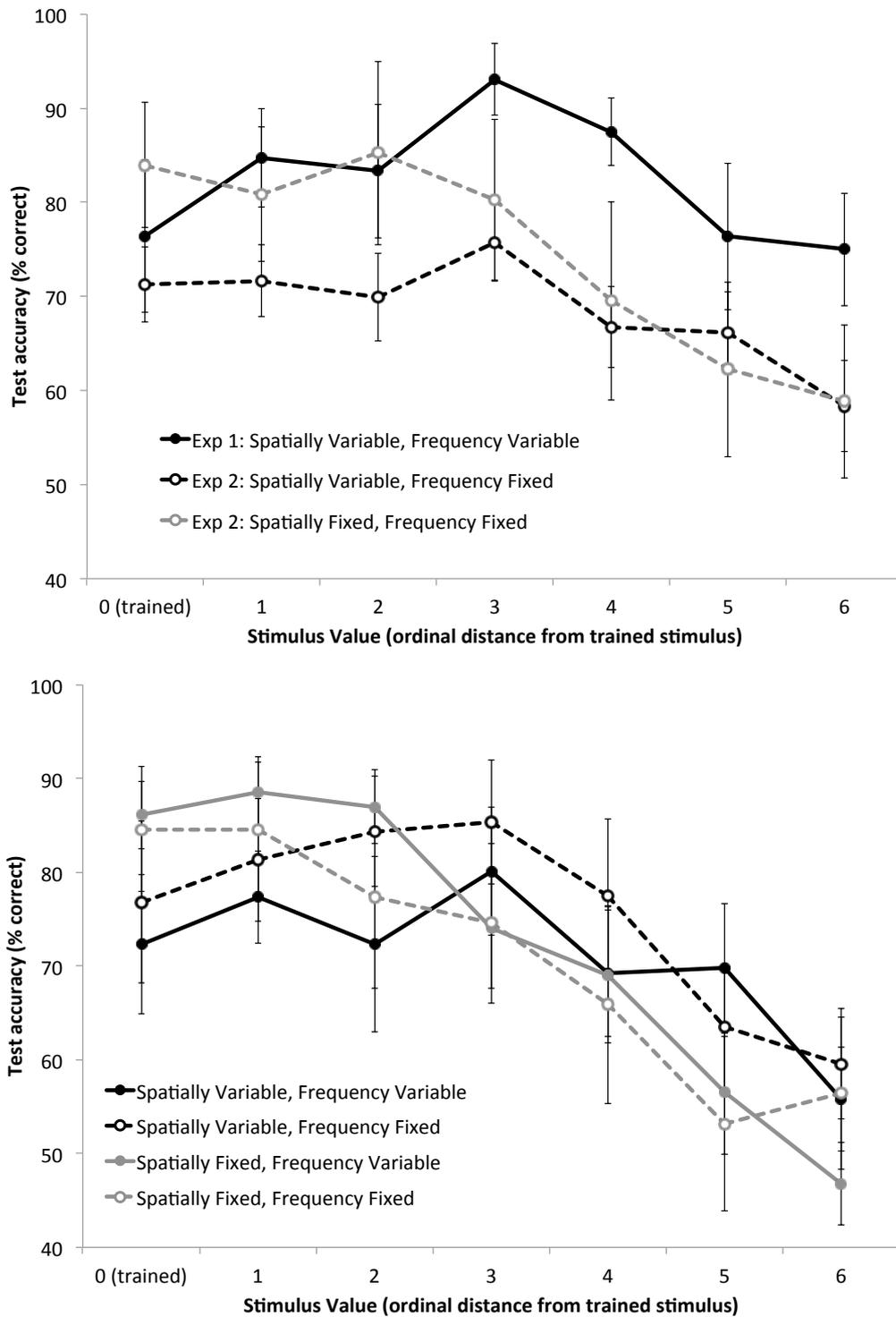


Figure 4. Mean test accuracy as a function of stimulus value, i.e. distance from the nearest training stimulus (S). Top panel shows test accuracy for Experiment 1 and the spatially fixed and variable groups in Experiment 2. Bottom panel shows test accuracy for Experiment 3. Error bars indicate SEM.

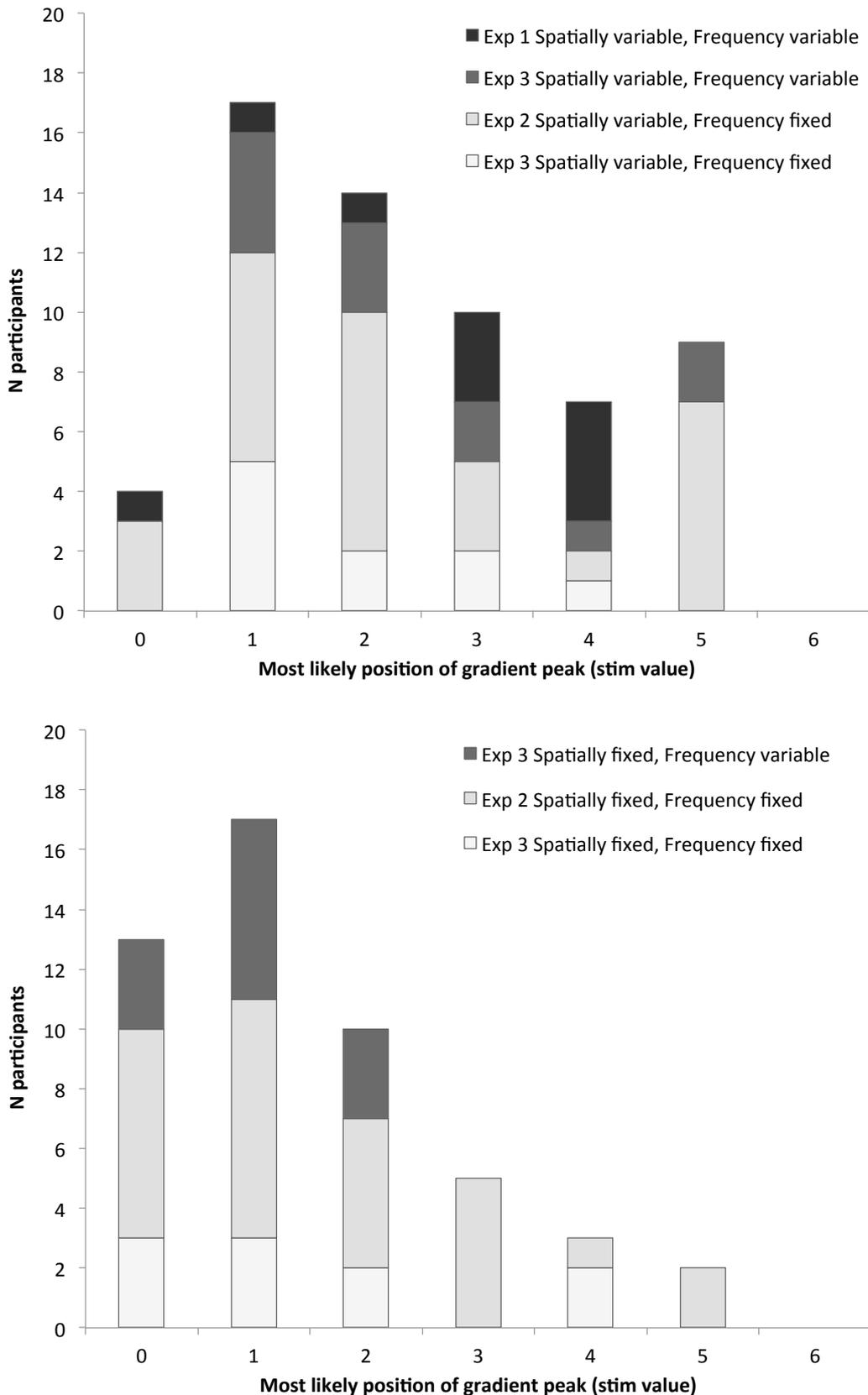


Figure 5. Number of participants (aggregated across experiments) for which peak accuracy (a) was most likely to fall at stimulus values 0 – 6. Top panel shows the four spatially variable conditions from Experiments 1-3. Bottom panel shows the three spatially fixed conditions from Experiments 2 & 3. Darker and lighter shades show frequency variable and frequency fixed conditions, respectively. Thirteen participants with essentially flat gradients were removed from this analysis (see text for details).

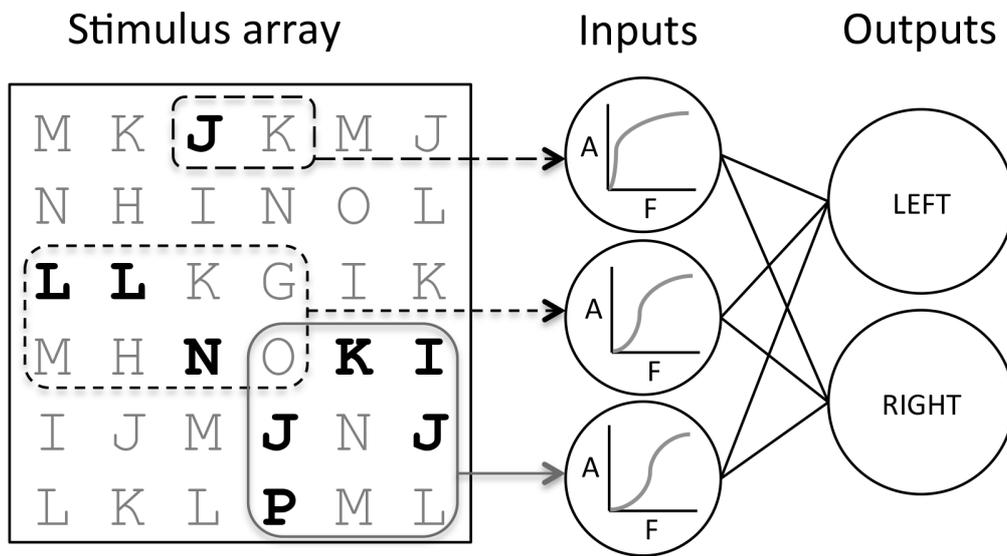


Figure 6. Schematic diagram showing the operations of the elemental associative learning model used to simulate post-discrimination generalization. Each representational element (labeled "Inputs") is sensitive to a randomly selected subset of the icons (e.g. those in bold in the figure) within a randomly determined region of the stimulus. These stimulus inputs have activation (A) as a function of the number of target features present (F) within this given region (with varying sensitivity as shown by the different sigmoid functions in the figure). Associative links between these Inputs and the Outputs used to generate a prediction are updated according to a prediction-error algorithm.

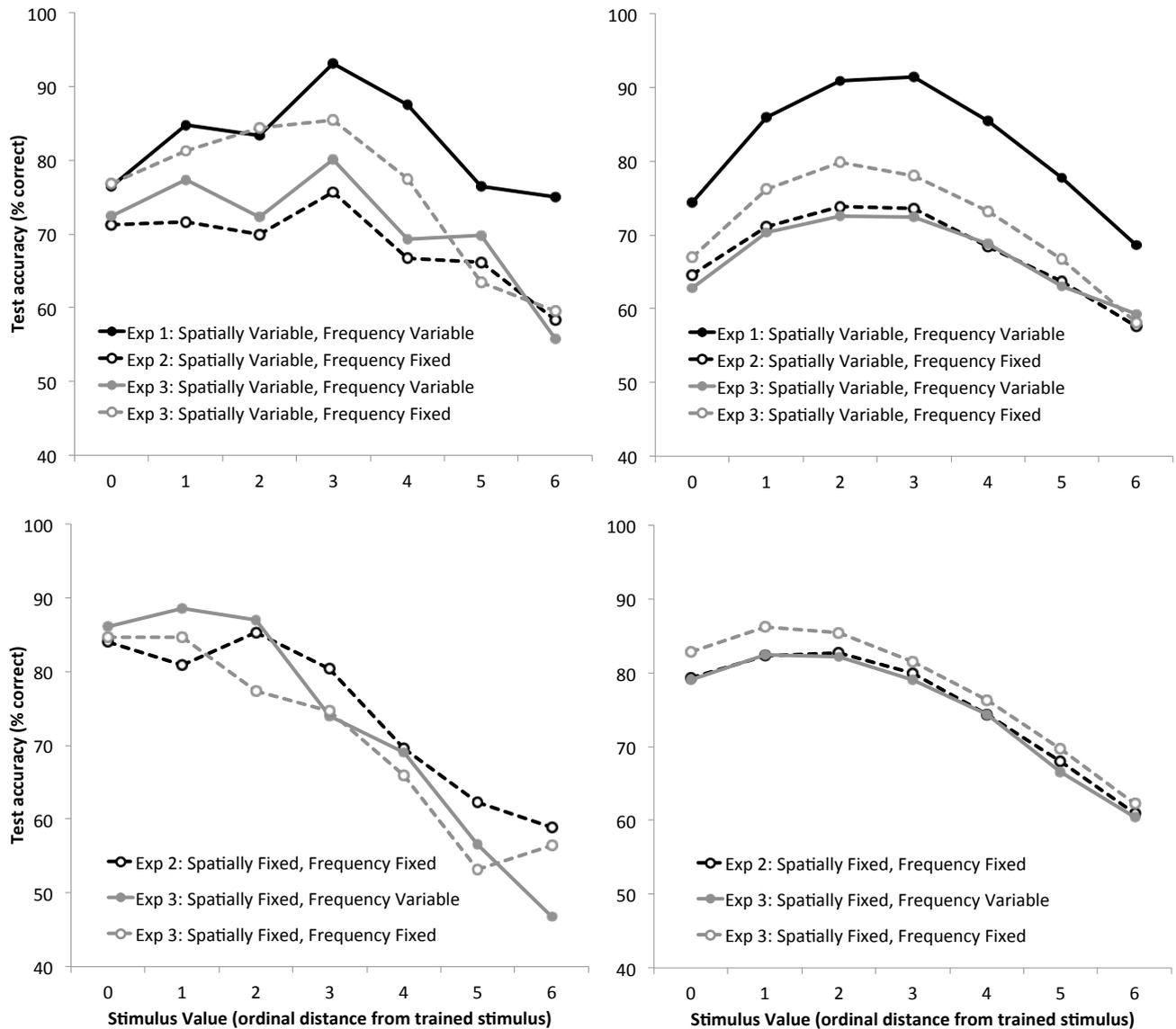


Figure 7. Left panels show generalization gradients from the three experiments grouped according to spatial variability (top panel: spatially variable; bottom panel: spatially fixed). Right panels show simulated data using an elemental model with simple stimulus sampling assumptions, again grouped according to spatial variability. All simulations were run with 100 representational elements, activated by 1-25 locations with the stimulus array and on average one third of the different types of icons, with a fixed learning rate $S = .01$, and a choice rule constant $k = 5.90$, $k = 2.87$, $k = 2.56$ for Experiments 1-3 respectively.