The Role of Similarity in Generalization

Matt Jones, W. Todd Maddox, and Bradley C. Love
[mattj,maddox,love]@psy.utexas.edu
University of Texas, Department of Psychology, 1 University Station A8000
Austin, TX 78712 USA

Abstract

Similarity is often regarded as a fundamental construct underlying stimulus generalization in category learning and many other domains. The key assumption of this approach is that multidimensional differences between stimuli are summarized by a single value before entering the decision process. The present study challenges this assumption by showing that category judgments depend on the full relationship between present and past stimuli, in a way that cannot be mediated by a unidimensional similarity measure. Approaches based on response generalization, knowledge partitioning, and distributional representations are also shown to be insufficient to account for our findings.

Introduction

Similarity has long been held to underlie a wide range of cognitive processes. Seminal work by Shepard (1957) showed that stimulus generalization in conditioning and identification tasks can be explained in terms of similarity between stimuli. This approach has since been extended to many other tasks, including categorization (Medin & Schaffer, 1978) and inductive reasoning (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). An important finding has been that similarity is not constant; rather, it changes systematically as a function of which stimulus attributes are relevant to the task (Heit & Rubinstein, 1994; Nosofsky, 1986). However, it is still generally assumed that similarity is well defined for any one judgment, context, and attentional set. This critical assumption holds not only for spatial models of similarity, but also for feature-set models (Tversky, 1977) and approaches based on internal relational structure (Markman & Gentner, 1993).

The present study challenges the assumption that generalization is based directly on similarity. We describe an experiment using a four-category classification task in which subjects must attend to two dimensions simultaneously, but must use these sources of information in different ways. The principal finding is that multiple generalization gradients are simultaneously active for different aspects of the category judgment. Thus performance in this task cannot be explained in terms of a single similarity function (even one that changes from trial to trial). We argue that the failing of the similarity approach is that it assumes the relationships between multidimensional stimuli are reduced to a single value before this information is passed to the decision process. That is, similarity acts as a mediator or sufficient statistic. Instead, it appears that people use the full multidimensional relationship between stimuli, and in particular the alignment between stimulus differences and category differences, in making category judgments.

Recency approach to generalization

Jones, Love, and Maddox (2006) demonstrate how stimulus generalization can be directly measured during a probabilistic classification task through analysis of decisional recency effects. Specifically, they found that responses are biased towards the feedback given on the previous trial, and that the strength of this bias is directly determined by the difference between present and previous stimuli. The effect of the previous feedback thus represents generalization from the previous stimulus to the current one.

Jones, Maddox, and Love (2005) found that when one stimulus dimension is predictive of the category label and another dimension is irrelevant, generalization becomes selectively dependent on the diagnostic dimension (Fig. 1). This finding is consistent with accounts of selective attention that assume similarity adapts to weight task-relevant dimensions more heavily (Kruschke, 1992; Nosofsky, 1986). In other words, stimuli differing along the diagnostic dimension become less similar than stimuli differing along the irrelevant dimension. This adaptation of generalization is directly observable through analysis of recency effects, as elaborated below.

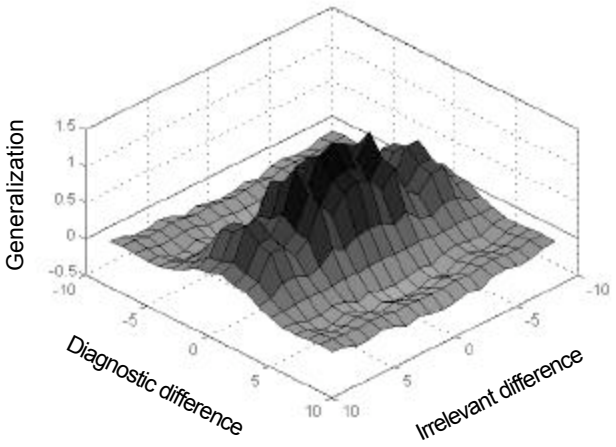

Figure 1: Selective generalization in a 2-category classification task (Jones et al., 2005, Expt. 1, Condition F). Horizontal axes indicate the difference between successive stimuli. Vertical axis shows strength of generalization, defined as the effect (in log-odds) of the previous feedback on the response to the present stimulus.

Empirical Investigation

The assumption that generalization is based on similarity was tested using a four-category probabilistic classification task. The structure of the categories used is illustrated in Figure 2. Stimuli in the task are Gabor patches, varying in frequency and orientation. Frequency is predictive of whether a stimulus lies in category A or C versus B or D, whereas orientation is predictive of A or B versus C or D. Therefore both dimensions are equally relevant to the task, but subjects' responses can be decomposed into components that, normatively, each depend on only one dimension. Each component is isomorphic to a two-category classification task with one diagnostic and one irrelevant dimension.

Data were analyzed according to this decomposition, and separate generalization gradients were obtained for each subtask, following the same approach as Jones et al. (2005). Importantly, the subtasks are merely constructs for the purposes of data analysis. On each trial the subject gives a single response from among the four categories, with responses for each subtask inferred at the time of analysis. Moreover, subjects were not given any instructions about the structure of the categories; they were merely told that there would be four categories for them to learn. In other words, the two subtasks are facets of the same judgment, and thus any similarity-based account must predict that they rely on the same similarity function. Therefore, if generalization is determined by similarity, then the generalization gradients from the two subtasks should be identical. However, if generalization is based on the full multidimensional relationship between present and past stimuli, then it might adapt in opposite directions for the two subtasks. We term this the *split-selective attention* effect, because it would indicate subjects are allocating their attention in different ways for different aspects of the task.

Method

Participants. Forty members of the University of Texas, Austin, participated for payment or course credit.

Stimuli. Stimuli were 6-cm square Gabor patterns (sine-wave gratings within a Gaussian envelope), varying in the frequency and orientation of the grating. The primary category structure involved 113 stimuli, arranged as shown in Figure 2. In addition, 13 extreme stimuli from each category, not pictured, were used during training.

Design. Every subject was tested on the same category structure (Fig. 2). The structure is fully probabilistic, such that every stimulus has a positive probability of occurring in any category. Outcome probabilities follow a logistic function along each dimension; for example, the probability that a stimulus lies in category C is given by $P[\mathrm{C}] = [(1 + e^{\sigma_{\mathrm{freq}}(S_{\mathrm{freq}} - \mu_{\mathrm{freq}})})(1 + e^{\sigma_{\mathrm{ori}}(S_{\mathrm{ori}} - \mu_{\mathrm{ori}})})]^{-1}$. Here S_{freq} and S_{ori} are the dimension values of the stimulus, μ_{freq} and μ_{ori} are the centers of the stimulus ranges, and σ_{freq} and σ_{ori} are constants set such that the maximal outcome probability for each category is 90%.

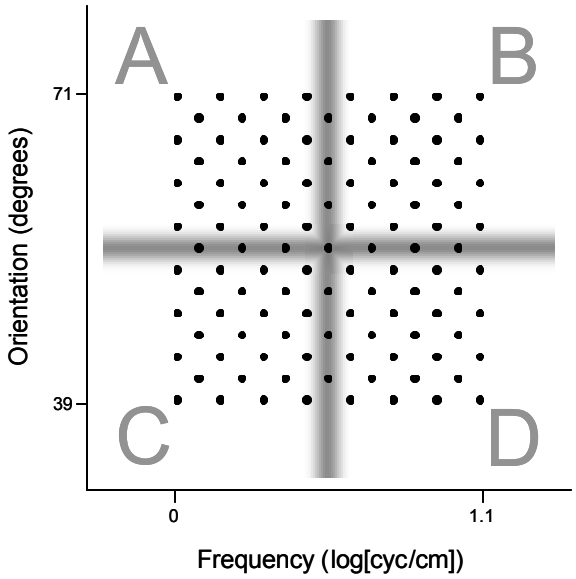

Figure 2: Experiment design. Dark circles indicate stimuli. Letters and blurred grey lines indicate category structure, although feedback is fully probabilistic.

Procedure. The experiment consisted of a training phase followed by a testing phase. The training phase used 13 extreme stimuli for each category and lasted 100 trials. Feedback during training was deterministic. This phase was necessary because piloting showed subjects perform very poorly on the probabilistic four-category task if they are not first taught the general arrangement of the categories. The testing phase used the 113 stimuli shown in Figure 2 and lasted 400 trials. Feedback during this phase was probabilistic, following the formula given above. At the start of the testing phase subjects were told that the categories were the same but that they would now be shown borderline items.

On each trial, a stimulus was randomly selected from the pool for the current phase and presented in the center of a 43-cm computer monitor on a black background. The subject then responded by pressing one of four keys on a keyboard. The word "Correct" or "Wrong," together with the correct category for that trial, was then presented below the stimulus for 1s. The monitor went blank for .5s before the start of the next trial.

Analysis

Responses and feedback from each trial were decomposed into two subtasks as shown in Table 1. Each subtask contains data from every trial but collapses the four categories to two. The effective categories are A∪C and B∪D in Subtask F and A∪B and C∪D in Subtask O. Therefore each subtask is logically identical to a two-category task with only one relevant dimension: frequency in Subtask F and orientation in Subtask O. The derived data for each subtask were analyzed using the sequential generalization model of Jones et al. (2005, 2006). Parameters obtained from fits of this model provide an estimate of the empirical generalization gradient.

Table 1: Decomposition of 4-category task into subtasks

	Category/Response			
Subtask	A	B	C	D
F	0	1	0	1
O	1	1	0	0

Notes: Entries indicate how each category is coded for each subtask. For Subtask F, only frequency is relevant; for Subtask O, only orientation is relevant.

The formal characterization of the sequential generalization model is as follows (for details and empirical validation, see Jones et al., 2006):

$$\text{logodds}(R_n) = F_{n-1}\Gamma(S_n, S_{n-1}) + \sum_i w_i (S_{n,i} - cS_{n-1,i}) + w_0. \quad (1)$$

This formula expresses the current response log-odds as a sum of short- and long-term contributions. The first term on the right side of Equation 1 represents generalization from the previous trial. The strength of generalization is given by Γ, which is a function of the present and previous stimuli S_n and S_{n-1}. The direction of the generalization effect is determined by the previous feedback F_{n-1}, which is coded here as ±1. Thus the present response tends towards the previous feedback to an extent determined by Γ. The remainder of Equation 1 represents the effect of long-term knowledge, which is included in the model to allow unbiased estimates of short-term generalization (Jones et al., 2006). Here $S_{n,i}$ represents the value of stimulus n on dimension i, w_i are association weights, and w_0 is an intercept or response bias term. The previous stimulus is included to model perceptual contrast effects, represented by c.

Two approaches are useful for estimating the generalization gradient Γ. First, Γ can be treated as a non-parametric function of the (vector) difference between S_n and S_{n-1}, by estimating a separate value for every possible difference. This approach yields a non-parametric mapping of the empirical generalization gradient (as in Fig. 1). The only assumption is that the two-dimensional gradient can be expressed as a product of gradients on each dimension (Nosofsky, 1986).

Second, Γ can be estimated from a parametric family. In the present study, parametric estimation of Γ follows previous research showing that generalization in category learning is best modeled by a Gaussian function of the distance between stimuli (Jones et al., 2005, 2006; Nosofsky, 1986). Therefore Γ is taken to be of the form

$$\Gamma(S_n, S_{n-1}) = m + ke^{-\sum \alpha_i (S_{n,i} - S_{n-1,i})^2}. \quad (2)$$

The intercept term m is included because of the finding of negative generalization between highly dissimilar stimuli (Jones et al., 2006). The α parameters determine the degree to which generalization depends on discrepancies along each dimension. Selective generalization corresponds to changes in α in response to the category structure, with larger values for more diagnostic dimensions (Jones et al., 2005). According to accounts of generalization based on similarity and selective attention, a large value of α represents increased attention to the corresponding dimension, which produces a decrease in similarity between stimuli differing on that dimension (Nosofsky, 1986).

To summarize, the sequential generalization model allows measurement of the pattern of generalization from the previous trial as a function of the relationship between present and previous stimuli. This is accomplished by assessing the effect of the previous feedback while controlling for the contribution of long-term knowledge. Comparison of the gradients obtained for the two subtasks of the present study provides a test of whether generalization was uniform across different components of the category judgment. This in turn tests the claim that generalization is based on similarity.

In all analyses, frequency and orientation were transformed to lie on a common scale, ranging from 1 to 15 in integer steps. All model fits are based on data from the testing phase only, and are based on maximum likelihood.

Results and discussion

The sequential generalization model (Eqs. 1 & 2) was applied to the derived data for each subtask, both for the group and for each subject. First, the nonparametric version of the model was applied to the group data to obtain nonparametric generalization gradients for each subtask. Long-term knowledge (w and c parameters) was allowed to vary among subjects. The gradients obtained for each subtask are displayed in Figure 3. As can be seen, the gradient for Subtask F is steeper along the frequency dimension than along the orientation dimension; the opposite pattern holds for Subtask O. Thus generalization for each subtask depends relatively more on the corresponding diagnostic dimension. To test the reliability of this difference, data from both subtasks were fit simultaneously, with the constraint that the two gradients were identical. The goodness of fit of this model was significantly worse than the combined fits of the previous models, $\chi^2(29) = 61.05$, $p < .001$. Therefore the generalization gradients differed between subtasks.

Second, the parametric version of the model (Eq. 2) was fit separately for each subject. To compare generalization between subtasks, a selective generalization measure was computed, separately for each subject and subtask, as $\beta = \alpha_{freq}/(\alpha_{freq} + \alpha_{ori})$. This variable measures the relative influence of the two dimensions in determining strength of generalization, and is constrained to lie between 0 and 1. The difference in β between the two subtasks is a measure of the split-selective attention effect.

Average values of β for each subtask, along with primary parameters from the long-term component of the model, are presented in Table 2. As can be seen, β is greater in Subtask F than Subtask O, again indicating that generalization for each subtask depends relatively more on the corresponding

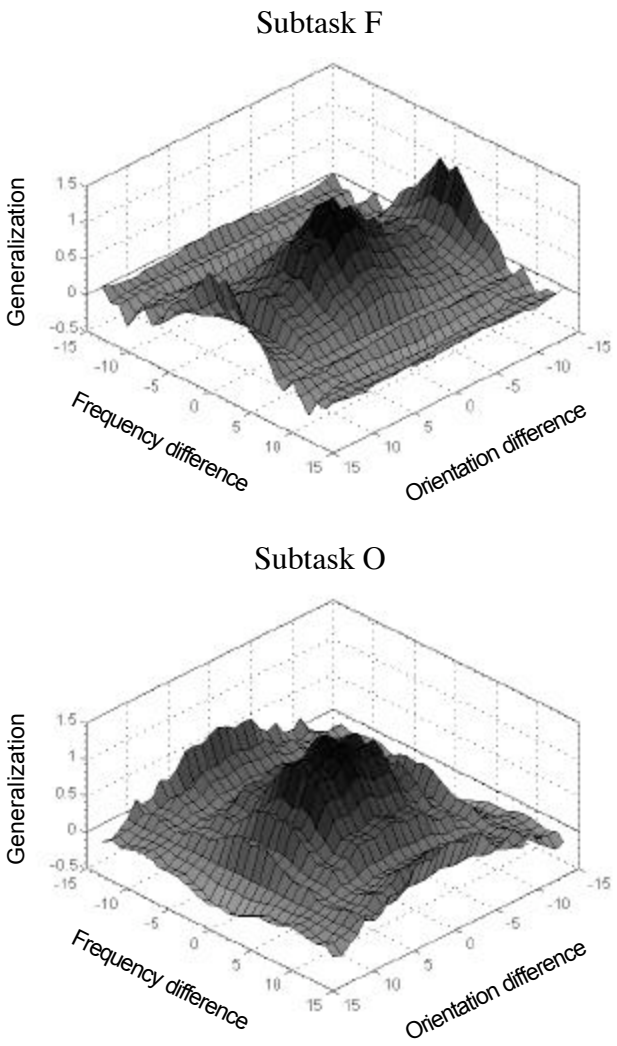

Figure 3: Non-parametric generalization gradients for each subtask. In both cases, generalization is weaker when successive stimuli differ on the diagnostic dimension (frequency for F, orientation for O) than when they differ along the irrelevant dimension.

diagnostic dimension.[1] This difference is significant by a paired-samples t-test, $t(39) = 1.80$, $p < .05$ (one-tailed). Furthermore, the strength of the split-selective effect is positively correlated to long-term knowledge of the category structure, defined as $w^{F}_{freq} + w^{O}_{ori}$ (with superscripts indicating subtask), $r = .417$, $p < .01$. Therefore the more subjects learned the category structure, the more they were able to differentially allocate their attention in the two subtasks.

[1]The fact that β is further from .5 in Subtask F than in Subtask O is merely a scaling effect – overall, generalization depends more on frequency than on orientation. This is also evident in the non-parametric gradients (Fig. 3). This observation and the fact that long-term cue use (w) was stronger for frequency in Subtask F than for orientation in Subtask O (see Table 2) suggest that frequency enjoys greater baseline salience for these stimuli.

Table 2: Primary measures from individual model fits

Subtask	β	w_{freq}	w_{ori}
F	.618	.299	.003
O	.473	.011	.195

Notes: β is selective generalization measure; greater values indicate more attention to frequency over orientation. w parameters measure long-term cue use.

Simulation

A series of simulations was conducted to test whether a similarity-based model can account for the pattern of generalization found in the present experiment. The simulations were based on ALCOVE, an influential model of category learning that has been used to explain a wide variety of classification phenomena (Kruschke, 1992). ALCOVE categorizes stimuli based on their similarity to exemplars stored in memory. Associations between stored exemplars and categories are updated by error-driven learning. This iterated updating produces recency effects, which are moderated by the similarity between successive stimuli (Jones & Sieck, 2003). That is, ALCOVE predicts similarity-based generalization from the previous trial. Furthermore, ALCOVE includes an attentional learning mechanism that modifies its similarity function, or generalization gradient, to improve performance. Thus ALCOVE is also able to explain the selective generalization effect found by Jones et al. (2005). ALCOVE therefore seems one of the most likely candidates to explain the split-selective attention effect from within the similarity framework.

Two versions of ALCOVE were simulated (see Fig. 4). The first is the standard version, in which each category is represented by a single output node. We refer to this version as the *unified model*. The second version, the *split-task model*, assumes that categories are explicitly represented in terms of the subtask decomposition used in the empirical analyses, with a pair of output nodes for each subtask. Response probabilities are calculated in the same manner as in the unified model (see Eq. 3 of Kruschke, 1992), but separately for each subtask. These values are then multiplied to obtain response probabilities for the overt categories (e.g., $P[A] = P[A \cup B] \cdot P[A \cup C]$).

The simulations showed that ALCOVE is unable to explain the split-selective attention effect. For all parameter values tested, the generalization gradients obtained from the two subtasks were statistically identical, with each depending equally on both dimensions. This is true even for the split-task version of the model, in which the classification response is explicitly generated from separate decisions on the two subtasks. The split-task version of ALCOVE fails to exhibit split-selective attention because decisions on the two subtasks still depend on the same attentional weights and hence the same similarity function. The activation of each hidden node, and hence the information passed to the output layer, only indicates the

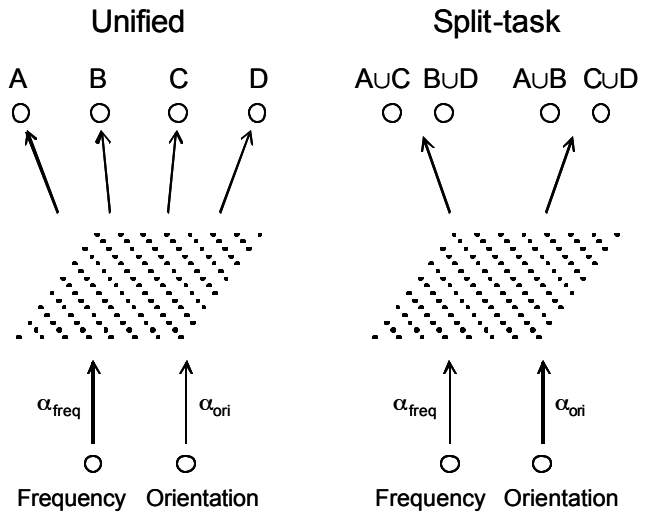

Figure 4: Illustrations of the two versions of ALCOVE used in simulations. See text for explanation.

similarity of that node's exemplar to the presented stimulus; it does not separately indicate their difference on each dimension. Prior to running the simulations it seemed possible that ALCOVE would exhibit split-selective attention via sequential effects from iterated learning, similar to the mechanisms by which it produces short-term generalization in the first place. However, this is not the case. Of course, our findings could be modeled by fully separating the processing for the two subtasks, by assuming two complete and independent copies of the model. However, allowing for separate similarity functions whenever an incompatibility arises, especially in such a post-hoc manner, undermines the predictive power of the similarity approach and renders it largely meaningless. Moreover, this approach abandons the assumption that relationships among stimuli are collapsed to a single similarity measure, and is more in line with our position that generalization is based on multidimensional information.

General Discussion

In one of the first empirical studies of categorization, Shepard, Hovland, and Jenkins (1961) investigated whether category learning can be explained by similarity-based generalization. Based on comparisons of error patterns between identification and categorization tasks, they concluded that it cannot. This conclusion was seemingly overturned by Nosofsky (1986), who showed how category learning can be well modeled by similarity-based generalization, given the additional assumption that similarity is systematically altered by selective attention. Using the sequential method for directly measuring generalization gradients in category learning, Jones et al. (2005) found that subjects learning different category structures exhibit different gradients, but again it could be assumed that this is due to shifts in attention leading subjects in different conditions to use different similarity metrics. Thus it could be argued that for each subject at each stage of learning, there exists a well-defined similarity metric underlying generalization.

The present study presents a much stronger challenge to similarity-based accounts of generalization, by demonstrating two different generalization gradients simultaneously active within the same judgment. The four-category classification task used here can be thought of as a superposition of two, two-category structures, each with a different relevant dimension (see Fig. 2). Just as was found when these structures were run separately, between subjects (Jones et al., 2005), generalization in each subtask was selectively more dependent on the relevant dimension. However, because in this study the two subtasks were in reality aspects of a single judgment, the differing gradients cannot be explained by a shift in similarity due to selective attention. This finding, termed the split-selective attention effect, demonstrates that the cognitive processes underlying generalization are more sophisticated than similarity accounts allow for. This conclusion is further supported by the simulations with ALCOVE, which is unable to exhibit split-selective attention.

An additional implication of this study is that subjects systematically generalize among categories. Extant models of category learning assume that observation of a stimulus in a given category is used as evidence regarding the membership of subsequent stimuli in that same category, but not as evidence about other categories (except indirectly, through response competition). However, in the present experiment it was seen that observation of a stimulus in one category can be taken as evidence in favor of other categories. For example, observation of a stimulus in category A led to an increased tendency to place the following stimulus in category B, provided the two stimuli were similar in orientation.

The idea that reinforcement of one response can increase the tendency for other responses is termed response generalization, and was predicted by Shepard's (1957) original generalization model. The present study demonstrates that response generalization is an important component of category learning. However, response generalization alone is not sufficient to explain our results. Shepard's model of stimulus and response generalization assumes that the two processes occur independently. Presentation of a stimulus activates knowledge about other stimuli based on their similarity, and the resulting response tendencies generalize to other responses, again based on similarity. Thus the degree to which observation of stimulus X lying in category A will be used as evidence for classifying stimulus Y into category B is a function of the similarities between X and Y and between A and B. In contrast, generalization in the present experiment was determined by the *correspondence* between the relationship between successive stimuli and the relationship between successive categories. Specifically, generalization is strong only when the dimensions on which the stimuli differ are the same as those on which the categories differ. Collapsing the multidimensional differences into unidimensional similarities before combining information about stimuli with

that about responses eliminates information about this critical correspondence.

Our proposal, then, is that generalization is based on alignment of stimulus differences with response differences, much like in analogy formation (Gentner, 1983). For each dimension, if the difference between present and previous stimuli is small, then categories are favored that are close to the previously reinforced category on that dimension. If the difference is large then categories that differ on that dimension are favored. This process is consistent with the pattern of generalization seen with unidimensional stimuli (Jones et al., 2006), and in that case is equivalent to an explanation based on similarity. However, the two explanations diverge in the multidimensional case, because similarity does not contain the information necessary to support generalization decisions on multiple dimensions simultaneously.

We suggest that the similarity approach has been successful to date because it was only tested in relatively simple tasks, generally involving only two categories. The present task goes beyond past research in that it includes multiple categories having different relationships to one another. Therefore the relevance of one stimulus' category membership to that of another is not a unitary proposition, but varies between the different aspects of the judgment. These relevancies cannot be summarized by any global similarity metric, but depend on the detailed, multidimensional relationship between the two stimuli.

Two other theoretical approaches deserve mention as they relate to split-selective attention. First, theories based on general recognition theory (Ashby & Townsend, 1986) assume that categories are represented in terms of their distributional properties, such as mean and variance on each dimension. If different categories are associated with different variance structures, then generalization of different category labels might be assumed to follow different gradients. However, categories in the present experiment all depended equally on both dimensions. Moreover, the pattern of generalization found did not vary according to individual categories but according to the relationship between pairs of categories (i.e., generalization between A & B and between C & D depended more on orientation, whereas generalization between A & C and between B & D depended more on frequency).

Second, Yang and Lewandowsky (2003) propose that people faced with a complex categorization task develop separate parcels of knowledge each applicable to a subset of the stimulus space. This strategy is referred to as knowledge partitioning. Knowledge partitioning can lead to more complex patterns of generalization than simpler similarity-based theories (e.g., Kruschke, 1992; Nosofsky, 1986), as attention might be allocated differently depending on which context is activated. However, knowledge partitioning cannot explain split-selective attention, because the phenomenon does not involve different generalization in different contexts, but rather different generalization for different aspects of the same judgment.

Acknowledgments

This work was supported by NRSA F32-MH068965 from the NIH to MJ, NIH Grant R01-MH59196 to WTM, and AFOSR Grant FA9550-04-1-0226 and NSF CAREER Grant 0349101 to BCL.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154-179.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 411- 422.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*, 316-332.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 1066-1071.

Jones, M. & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 626-640.

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

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32*, 517-535.

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

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

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185-200.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325-345.

Shepard, R., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75,* 1-42.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-52.

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