Alternation as a Relational Category

**Anonymous CogSci submission**

**Abstract**

Key questions in the study of categorization are how individuals form categories from experience and extend that knowledge to assess membership of novel examples. Popular accounts predict generalization to be based on either similarity to reference points or the application of rules or bounds. However, recent data show that some categorization behavior defies the predictions of leading accounts. Expanding on these findings, in the present experiment participants learned a one or two dimensional alternating category structure and were then tested on near and far transfer tasks. Findings reveal that individuals can extend a learned alternating category structure across multidimensional spaces and increasingly distant generalization regions. A new far transfer task suggests that alternation is learned in the form of a relational category at the level of the domain space within which subjects have learned an attribute-based categorization scheme operating at the item level.

**Keywords:** Categorization; Classification Learning; Relational Categories; Generalization of Learning

# Introduction

The way in which individuals determine the category membership of new items is a question of persistent and profound interest in the field of categorization. Individuals may decide by comparing new items to previously stored category members (e.g., Nosofsky, 1984), to a representation of the average category member (drawing on prototype theory, e.g., Rosch & Mervis, 1975), or by employing a rule that explicitly separates items into different categories.However, there is recent evidence that not all generalization behavior can be comprehensively captured by currently held theories of categorization.  
 Kurtz and Wetzel (2021) trained participants on a one-dimensional alternating category structure (e.g., A B A B) and found that a majority of participants generalized new test items according to the overarching alternating pattern during an unsupervised test phase and only a minority of participants generalized based on proximity or similarity. Specifically, the untrained region right next to the training region represented a critical test: items in this space are quite close and similar to the B items at the extreme of the alternated training set, but they would be considered A items if extrapolating the alternation pattern into the untrained region. In a second experiment testing an alternating structure based on one diagnostic dimension embedded in a two-dimensional domain, the researchers found a decrease in the proportion of participants that generalized the alternation, but still a notable level and clearly more than would be expected from chance. These findings indicate that some generalization strategies are not captured by proximity based descriptions of generalization.  
 A potential non-proximity, or similarity, based description may lie in the relational category literature. Members of relational categories do not necessarily have feature-based similarity, but are instead unified by fulfilling the same core relationship – having relational similarity (Gentner, 1983; Gentner & Kurtz, 2005; Gentner & Markman, 1995). Since relational category members need not share feature-based similarity, a relational account is proposed as a contending explanation of the generalization behavior observed in Kurtz and Wetzel (2021). This is intriguing and novel territory in that that relational category does not take individual items in the categorization domain as examples; instead the totality of the categorization domain represents a singular example of the relational category of alternating systems (that includes checkerboards, striped patterns, the day/night cycle, opportunities for competitors to score points in games, etc.).  
 The present investigation seeks to expand on the findings of Kurtz and Wetzel (2021) by replicating the single dimension findings and expanding on the two-dimensional findings with a modified two-dimensional structure in which both dimensions are category-relevant. We created a two dimensional category structure with a partially diagnostic non-alternating dimension, and a fully diagnostic alternating dimension. This was intended to evaluate whether subjects remain sensitive to the alternating dimension when given a non-alternating dimension that can inform category membership. This category structure also allowed us to evaluate whether the alternating structure would be generalized to regions which would not explicitly be continuations of the alternating pattern of the training items. Further, in order to test the potential applicability of relational categories, a transfer task was administered to evaluate the proportion of alternating responses in an unsupervised environment with stimuli visually entirely distinct from that used in training and near transfer. Given the importance of relational similarity in the accurate application of analogies during far (superficially distinct) transfer tasks (Gentner & Kurtz, 2005; Gentner & Markman, 1995), evidence of successful far transfer from individuals who learned the alternating category structure would suggest that alternation-based generalization behavior may be best understood in terms of activation and application of a relational category.

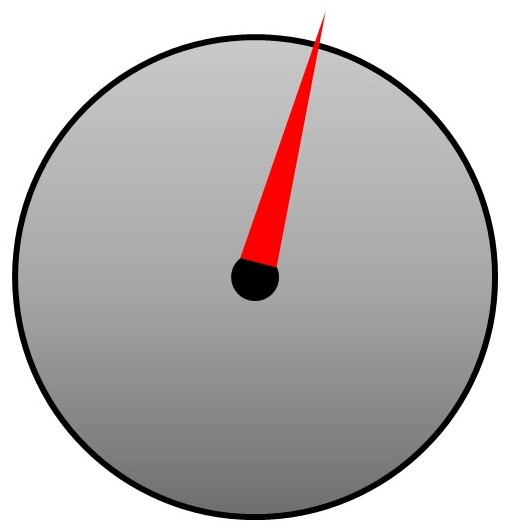
# Method

## Subjects

102 undergraduates participated in the experiment for partial fulfillment of a participation requirement for a psychology course.

## Stimuli

Categorization Task. There were two sets of stimuli for the categorization task corresponding to two between-subjects conditions. One stimuli set varied along a single dimension (1D condition), while the other stimuli set varied along two dimensions (2D condition).The one dimensional stimuli consisted of a gray circle with a red line anchored to the center of the circle and protruding outward at a certain angle (Figure 1).The angle of this red line in relation to the circle was the dimension that was varied. These 1D stimuli followed an alternating category structure, such that as the values along the variable dimension increased, the category labels corresponding to those feature values alternated

Figure 1: Sample Stimulus.

systematically. Training stimuli ranged from a line angle of 20 degrees to 179.5 degrees. Within this range, category exemplars were clustered in groups of three. Within a given cluster, items were all of the same category and were buffered by 5 degrees of angle difference. Between two consecutive clusters, there was an angle difference of 51.5 degrees. The alternation of category labels happen at the level of the cluster (all items within a cluster were of the same category, while all items of the immediately preceding cluster were of the contrasting category). The line angle of the near-transfer test items ranged from 226 degrees to 334 degrees distributed in a pattern in keeping with that of the training items (see Figure 2).

The two-dimensional stimuli were of the same type as shown in Figure 1; however, these stimuli varied both in terms of line angle and in terms of circle diameter. These stimuli alternated with respect to category labels along the angle dimension (similar to the 1D stimuli), and were partially distinguishable based on the diameter – if classification judgments were made based only on diameter, then accuracy would be 50% (see Figure 3).

Figure 2: Representation of the single dimension category space. Stimuli had consistent diameter but varied in line angle. “A” and “B” indicate training items that are “Alpha” or “Beta,” respectively. Each “?” refers to a stimulus item shown during the test phase.

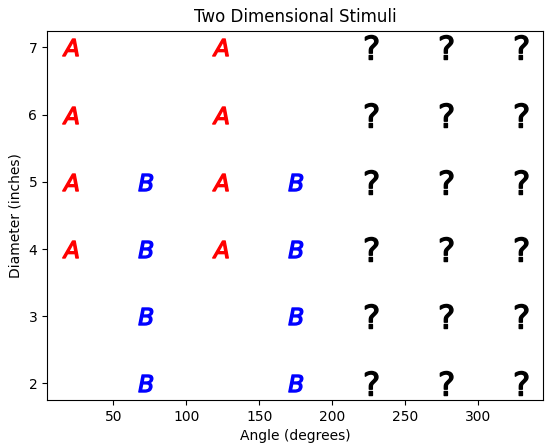
### Far Transfer Task. Eight marbles differing only in color (four red and four blue) were displayed in a random initial

### configuration prior to being arranged by the participant.

## Procedure

Participants were randomly assigned to one of three groups: The single dimension condition (1D, n = 35), two dimension condition (2D, n = 36), or control (n = 31). Control participants only experienced the far transfer task.

**Training Phase.** Participants were told they would be trained to evaluate “instrument readouts.” Stimuli were

Figure 3: Representation of the two dimension category space. Stimuli varied in diameter and line angle. “A” and “B” indicate training items that are “Alpha” or “Beta” respectively. Each “?” refers to a stimulus item shown during the test phase.

displayed individually on each trial. Participants were instructed to indicate category membership by selecting the “Alpha” or “Beta” button. Corrective feedback was provided after each response. During a training block each stimulus was evaluated once, in random order, for a maximum of ten blocks. Participants progressed to the near transfer component of the testing phase after correctly evaluating twelve consecutive training stimuli. Individuals who did not meet this criterion were removed from analyses.

**Near Transfer Phase.** In random order, a single training or test stimulus was shown, and participants indicated category membership by selecting the “Alpha” or “Beta” button. Each item was evaluated once and no feedback was provided. Upon completion, participants continued on to the far transfer phase.

Amount of alternation extension was operationalized via *testing regions*. The right half of each categorization space comprises the entire testing region for each structure (see Figures 2 and 3). Each third of the entire testing space is its own *test region*, with the region closest to the training stimuli designated the *first testing region*.

**Far Transfer Phase.** We note again that this component was the entire experiment for Control group participants. Experimental participants were informed that they were moving on to the next part of the experiment. This was somewhat ambiguous since the participants were completing multiple separate studies within a one-hour experimental session; therefore it could have been viewed on one extreme as an extension of the learning task with the instrument panels or on the other extreme as entirely unrelated. Participants were shown eight marbles (four red and four blue) in a random arrangement. Horizontally, along the bottom of the screen, there were eight empty boxes. Participants were given the following instructions “These marbles are currently in random positions. Please arrange them along a line into an ordering of your choice. Press the finished button when you are done.” Participants received no feedback.A successful alternation outcome on the far transfer test was operationalized as comprehensively and consistently alternating across the sequence of marbles either by 1’s (rbrbrbrb), by 2’s (rrbbrrbb), or by flanking (rrbbbbrr). A successful alternation was awarded a score of “1”while any other arrangement was assigned a value of “0.”

# Results

9 subjects were removed from the one dimension (1D) condition and 3 subjects were removed from the two- dimension (2D) condition for failure to meet learning criteria during training. All analyses were performed on participants that met learning criteria during training (1D: n = 26; 2D: n = 33).

Participants were considered proximity classifiers if ⅔ of items in the first testing region extended the value of the adjacent training region. Participants were considered to have extended the alternation pattern into the testing region *once* if ⅔ of items in the testing region closest to the training region extended the alternation pattern (Alpha, Beta, Alpha, etc.). Participants were considered to have extended the alternation pattern throughout the *entire* testing region if ⅔ of the items in each group of test stimuli properly extended the alternation pattern. This ⅔ rule was decided so that participants who were off by a single item classification per group of testing items would remain in consideration. To keep the one and two-dimension conditions consistent the ⅔ rule was applied in both, thereby allowing for a 2 item buffer in the two-dimension condition.

42.3% of 1D subjects and 33.3% of 2D subjects extended the alternation pattern into at least the first testing region. 27% of all 1D participants, and 33.3% of all 2D participants extended the alternation pattern throughout the entire testing region (aggregate near transfer categorization decisions for each condition visualized in Figures 4 and 5).

11.5% of 1D subjects classified items in the first testing region according to the closest training region (proximity classifiers). No 2D participants were considered proximity classifiers.

3.8% of 1D subjects split the entire category space such that most training items were designated Alpha and most test items were designated Beta. 6.1% of 2D participants split the categorization space similarly but designated most training items as Beta and most test items as Alpha. 12.1% of 2D subjects split the entire categorization space such that the upper half of the space was designated Alpha while the lower half was designated Beta.

In the far transfer task of arranging the marbles, 57.7% of 1D participants and 60.6% of 2D participants produced an alternation outcome. A majority of successful transfer responses alternated the sequence of marbles by 1’s (rbrbrbrb; 1D: n = 13; 2D: n = 19); very few alternated by 2’s (rrbbrrbb; 1D: n = 1; 2D: n = 0) or by flanking (rrbbbbrr; 1D: n = 1; 2D: n =1). Far transfer performance was evaluated via goodness of fit tests using the proportion of successful vs. unsuccessful transfer in the control condition as expected values. Significant differences in the proportion of successful transfer were found between the 1D condition (n = 26, M = 0.58) and control (n = 31, M = 0.36; *X2*(1, *N* = 57) = 5.51, p < .05, w = 0.31), and between the 2D condition (n = 33, M = 0.61) and control (*X2*(1, *N* = 64) = 9.81, p < .01, w = 0.39; Figure 6). Phi (w) indicates a medium effect size for each.

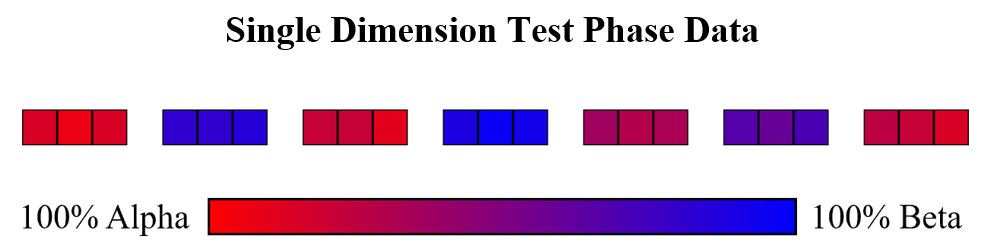


Figure 4: All 1D near transfer responses (n = 26). A single stimulus item is represented by a single square; color represents the proportion of participants that selected Alpha or Beta. The three rightmost groups indicate the three near transfer testing regions.

# 

Figure 5: All 2D near transfer responses (n = 33). A single stimulus item is represented by a single square; color represents the proportion of participants that selected Alpha or Beta. The three rightmost groups indicate the three near transfer testing regions.

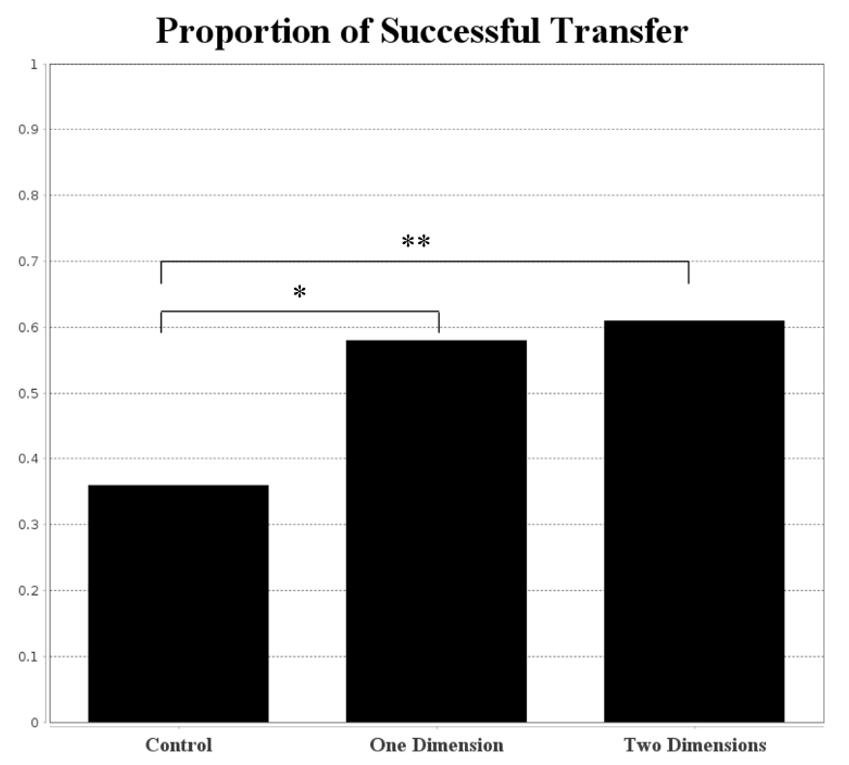


Figure 6: Proportion of alternation outcomes observed for each condition in the far transfer task.

## Simulation

As a possible explanation of the learning and generalization of alternating category structures, Kurtz and Wetzel (2021) report a simple neural-network model that simulates the qualitative patterns of human performance on a 1D alternating category structure; which, may suggest that the ability to learn and generalize an alternating category structure can emerge as simple inputs progress through a connectionist system.

Employing the connectionist principle of error driven learning (Rumelhart, Hinton, & Williams, 1986) and recreating the architecture described in Kurtz and Wetzel (2021), we constructed a simple neural network model to see if such a model could simulate the current behavioral observations. The model can be described as a multilayer perceptron (MLP): its architecture consisted of an input layer, a hidden layer with a single input node, and an output layer with a single output node. At the hidden node, a sine function was used, and at the output node a linear activation function was used. Kurtz and Wetzel found that a periodic activation such as a sine function was able to simulate various forms of alternation behavior. The model had two free parameters: learning rate (0.1) and initial weight range (a range from 6.0 to 6.1). The model was trained on data that represented alternating structures incorporating either a single dimension or two dimensions; and was trained for 2000 epochs with random item presentation. The data used as input was scaled between 0 and 1, and was of the same structure types as those given to human subjects. The model was not quantitatively fit to the human behavioral data, rather the model was trained on an alternating structure to see if it could produce various human-like qualitative outcomes. In the 1D case it was found that the model produced response probabilities that would appropriately predict item labels in the case of alternation, a unidimensional rule, and a partial alternation (Figure 7). However, when the input incorporates two features (such as our 2D condition), the model poorly simulates the human capacity to learn an alternating category (Figure 8). The model is able to predict item labels appropriately for the case of a unidimensional rule and partial alternation, but in the case of the full alternation the model fails.

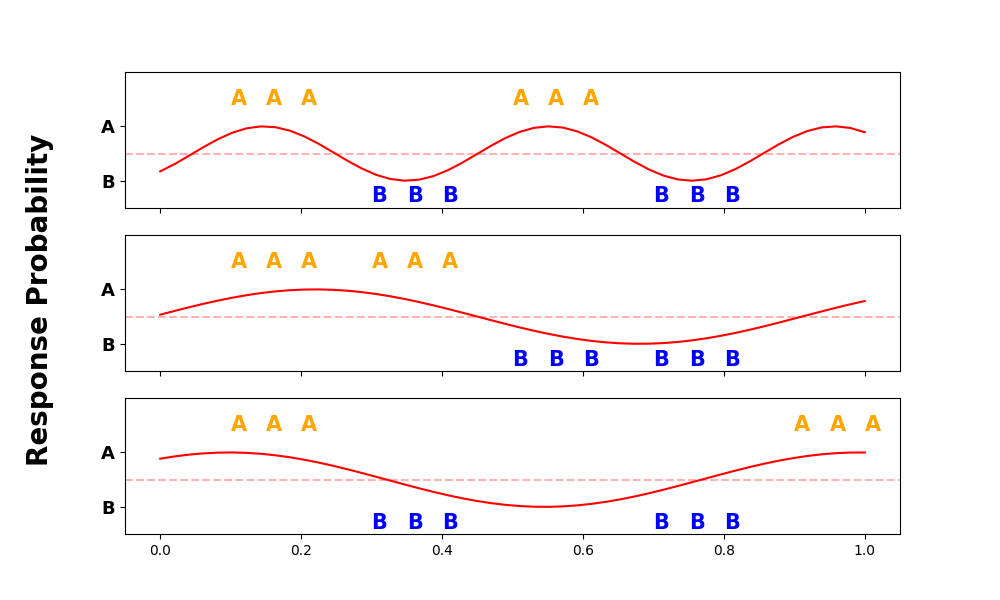


Figure 7: Given one dimensional input, an MLP using a sine function as hidden node activation is able to simulate various behaviors exhibited by humans when faced with an alternating category structure. The red line represents response probability along the single alternating dimension. The x-axis corresponds to the feature space, and the y-axis corresponds to response probability.

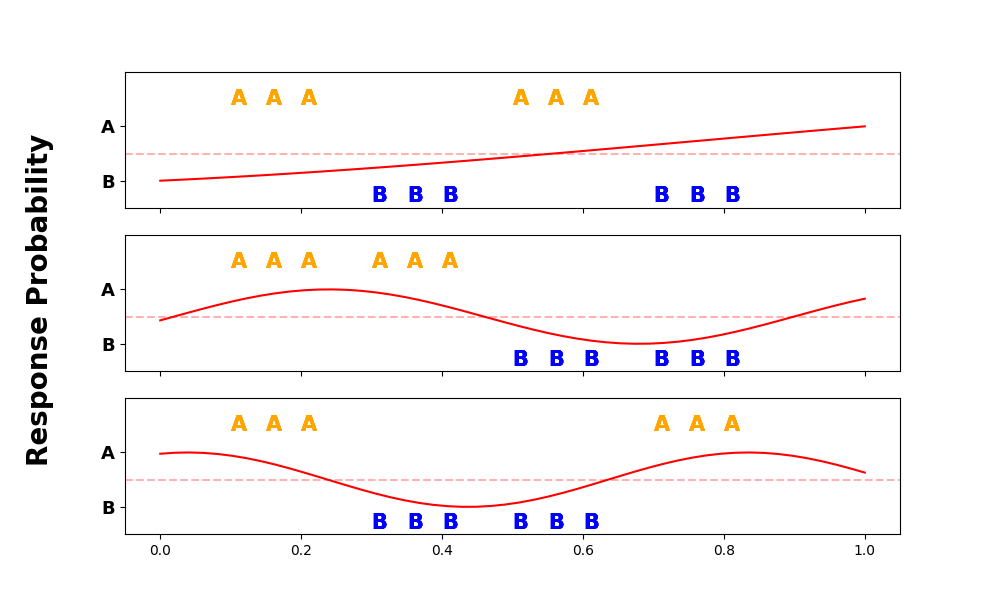


Figure 8: Given two dimensional input, an MLP using a sine function as hidden node activation fails to simulate the human capacity to learn an alternating category structure. The red line represents response probability along the alternating dimension. Input data incorporated two dimensions, but response probabilities are illustrated with reference to just the alternating dimension for simplicity and comparison to the one dimensional case. The x-axis corresponds to the feature space, and the y-axis corresponds to response probability.

# Discussion

**Near Transfer**

**One Dimension.** Kurtz and Wetzel (2021) found that 67.8% of their participants extended the alternation into the first testing region while 25.0% of participants extended the alternation throughout all three testing regions. In the present study a smaller proportion of 1D participants extended the alternation pattern into the first region (34.8%), but a similar proportion extended the pattern throughout all three testing regions (26.9%). Additionally, more proximity classifiers were observed (11.5%) than previously reported (Kurtz & Wetzel, 2021; 3.5%) on single dimension alternating stimuli.

The current study used a learning criteria of 12 consecutive items correct at test, and further used a 2/3 correct criteria for each group of test stimuli (e.g., if 2/3 of responses in the first testing region were consistent with alternation, then that participant was considered to alternate to the first testing region). However, Kurtz and Wetzel (2021) used more stringent criteria – 90% accuracy on training items during the testing phase and 100% of testing region responses needed to be consistent with alternation responses for a subject to be considered as extending the alternation. When applying this criteria, a similar proportion of participants were observed extending the alternation at least once (69.2%) and throughout the entire testing region (30.8%). There were still more proximity classifiers than previously reported (15.4%). However, after applying the stricter criteria, there were fewer subjects in consideration (1D: n = 13 as opposed to n = 26).

The exact same one-dimensional structure, stimuli, and feature values from Kurtz and Wetzel (2021) were used in our 1D condition. Given the similar proportion of alternation behaviors observed between the two studies (after employing the same learning and accuracy criteria), the one-dimensional findings of Kurtz and Wetzel (2021) were successfully replicated.

**Two-Dimension.** When testing the alternating structure on two-dimensional stimuli, Kurtz and Wetzel (2021) found that 34.8% of participants extended the alternation (only one testing region used). The present study observed a similar proportion of alternation extension (33.3% of 2D participants extended alternation into at least the first region). While a similar rate of extension was observed, it is important to note that the two-dimensional space employed by Kurtz and Wetzel (2021) had only one diagnostic dimension; the two-dimensional space used here had one diagnostic and one partially diagnostic dimension. Building off the single testing region findings of Kurtz and Wetzel (2021), we used three testing regions and observed that 27.0% of 2D participants extended the alternation throughout all three testing regions. No 2D participants categorized such that a majority of final training region responses matched the majority of responses in the first testing region (i.e., no proximity classifiers). This is much less than previously reported (Kurtz & Wetzel, 2021; 26.1%) for two dimensional alternating stimuli.

After applying the more stringent criteria of Kurtz and Wetzel (2021), a higher proportion of participants were observed extending the alternation at least once (75.0%), while 25.0% of participants were observed extending the alternation throughout the entire testing region. However, this drastically reduced the number of subjects considered in the 2D sample (n = 12, as opposed to n = 33).

The two dimensional category space used was arranged such that only a middle portion of the space alternated while very high diameter values were diagnostic of Alpha items and very low diameter values were diagnostic of Beta items (Figure 3). No 2D participants alternated items with middle diameter values while simultaneously maintaining the diagnostic separation of high and low diameter values. However, 12.1% of 2D participants split the entire category space such that the upper half of training and test items (higher diameter values) were mostly designated Alpha and the lower half of all items were mostly designated Beta (one subject example shown in Figure 9).

Overall, it has been shown that a reasonable proportion of individuals will extend an alternation pattern in both a one and two dimensional category space. This generalization pattern is not explained by proximity-based theories of categorization and therefore remains a point of considerable interest.

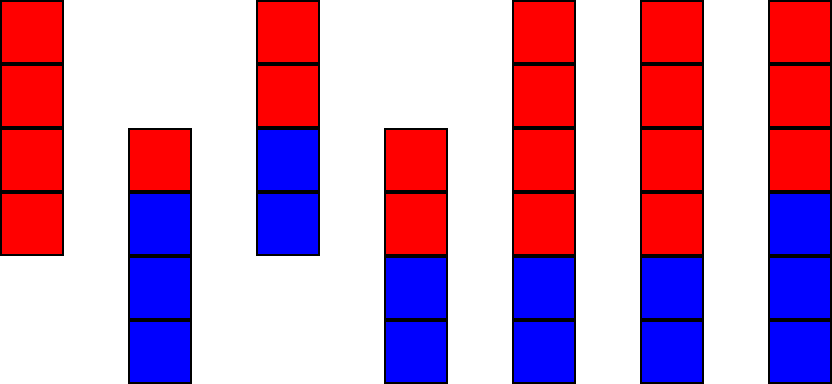


Figure 9: Single 2D subject who failed to extend the alternation during near transfer but did split the category space based on diagnosticity of circle diameter. Each square represents a single stimulus item. The three rightmost regions indicate the three testing regions.

**Far Transfer**

Since generalization via an alternating pattern of category membership does not depend on proximity or feature-based similarity, a far transfer task was used to evaluate whether alternation as a relational category could explain the generalization. Given the importance of structural alignment of relational information during analogical transfer (Gentner, 1983; Gentner & Kurtz, 2005; Gentner & Markman, 1995) it was reasoned that successful transfer of the alternation pattern from the training/testing stimuli to the marbles would indicate that alternation was at least supported by relational knowledge of the structure. It was found that a much higher proportion of individuals in the experimental conditions produced a successful alternation of the marbles compared to the control subjects. This supports the notion that a relational category may underlie the observed extension of alternating category membership.

**General Discussion**

The current work replicates and extends the findings of Kurtz and Wetzel (2021). Exemplar theory, prototype theory, and rule based theories all fail to predict the ability to learn and generalize an alternating category structure. Kurtz and Wetzel (2021) offered two possible explanations for which the current results provide important evaluation.  
The first possible explanation was that this complicated pattern of categorization could emerge as simple neural-level data is propagated through a connectionist system. When modeling one dimensional stimuli, this explanation is plausible. However, when two dimensions are incorporated this explanation (at least, for the proposed architecture) fails.

The second possible explanation was that subjects may accomplish the learning and generalization of an alternating category structure through relational reasoning. This would mean that rather than learning something just about the specific attributes of trained items, subjects are abstracting the relational category of alternation from the mapping between feature inputs and category labels. This explanation is lent support by far transfer task data: subjects given classification training were significantly more likely to invoke alternation during the far transfer task than were subjects who did not receive classification training. This strongly suggests either transfer or priming of an activated relational category given that there was no similarity between the two tasks.

# References

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.<http://dx.doi.org/10.1207/s15516709cog0702_3>

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside of the lab: Festschrift in honor of Douglas L. Medin* (pp. 151–175). Washington, DC: American Psychological Association.<http://dx.doi.org/10.1037/11156-009>

Gentner, D., & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (Ed.), *Similarity in language, thought, and perception* (pp. 111–147). Brussels, Belgium: BREPOLS.

Kurtz, J. K., & Wetzel, T. M. (2021). On the generalization of simple alternating category structures. *Cognitive Science, 45*.<https://doi.org/10.1111/cogs.12972>

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104–114.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.