# Models of Human Category Learning: Do They Generalize?

**Nolan Conaway (nconawa1@binghamton.edu)**
**Kenneth J. Kurtz (kkurtz@binghamton.edu)**
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

**Abstract**

Generalization to new examples is an essential aspect of categorization. However, recent category learning research has not focused on how people generalize their category knowledge. Taking generalization to be a critical basis for evaluating formal models of category learning, we employed a 'minimal case' approach to begin a systematic investigation of generalization. Human participants received supervised training on a two-way artificial classification task based on two dimensions that were each perfect predictors. Learners were then asked to classify new examples sampled from the stimulus space. Most participants based their judgments on one or the other dimension. Varying the relative levels of dimension salience influenced generalization outcomes, but varying category size (2, 4, or 8 items) did not. We fit two theoretically distinct similarity-based models (ALCOVE and DIVA) to aggregate learning data and tested on the generalization set. Both models could explain important aspects of human performance, but DIVA produced a superior overall account.

**Keywords:** generalization; categorization; formal models of category learning; similarity; cognitive modeling.

## Introduction

Categorization is an essential cognitive function – categories serve to organize knowledge and, critically, as a basis for extending knowledge to make sense of new experience. A full understanding of human categorization depends on developing models and theories that account for systematic patterns of human learning and generalization performance (for an overview of generalization, see Levering & Kurtz, 2010).

In classic research, Roger Shepard (1957, 1987) put forth the idea of a universal law in which stimulus generalization follows an exponential function of distance in psychological space. This work has had broad implications for theoretical models of categorization. Highly influential reference point models (such as the exemplar view) compute classification in a manner that closely follows Shepard's proposal. Specifically, the class membership of a known item is likely to be generalized to a new item if the two items are highly similar. The key additional design feature needed to account for human classification performance is the inclusion of a selective attention mechanism such that particular dimensions can matter more or less in the computation of similarity. Generalization performance (classification of previously unseen items) has been one of the most important important testing grounds in the debate between exemplar- and prototype-based accounts (e.g., Homa, 1984; Nosofsky, 1992; see also Medin & Schaffer, 1978 and the ensuing literature on behavioral experimentation and model-fitting with the 5-4 classification problem).

In a somewhat different approach to studying the generalization of category knowledge, researchers have investigated whether exemplar models can account for rule-like generalization after category learning (Erikson & Kruschke, 1998, 2002; Nosofsky & Johansen, 2000). In these studies, participants were asked to classify novel instances after learning an artificial two-way classification based on a unidimensional rule with exceptions. The critical test items were highly similar to the exceptions, but clearly classifiable using the rule. The outcomes of these studies were somewhat mixed and appear to depend on stimulus attributes and also on the structure of the categories that are learned.

The goal of the present research is two-fold: 1) to explore a different approach to investigating the psychology of category generalization; and 2) to use generalization performance as a basis to compare and differentiate models that are highly successful in fitting human learning data. Toward the first goal, our experimental approach is broadly comparable to the psychological studies of generalization discussed above: after a learning phase, participants are asked to classify novel examples. However, our work differs in that we use minimal category learning conditions (small numbers of examples that are readily assigned to two fully coherent classes). Our primary aim is to identify basic, systematic properties of generalization performance.

Regarding the second goal, the field presently offers a small group of formal models of category learning that are general purpose (applicable to any classification problem), that provide explanation at the level of process/mechanism, and that yield good fits to established benchmarks for human category learning. Within the realm of fitting human classification learning performance, there is some sense of having hit the ceiling in terms of differentiating among these models despite their having distinct explanatory elements. Our rationale is that models that do quite well in fitting learning data may diverge in their ability to account for patterns of generalization performance. In particular we are compelled by the prospect of fitting model parameters to the learning data and then holding the models to these values in evaluating ensuing generalization (as discussed below). Toward this end, we evaluate two successful models: a canonical representative of the reference point approach, ALCOVE (Kruschke, 1992) and an updated

version of a competing theoretical alternative, DIVA (Kurtz, 2007).

**ALCOVE**. ALCOVE is an exemplar based adaptive network model. According to the model, categories are represented by individual exemplars stored in memory. ALCOVE learns to classify by adjusting association weights between exemplar nodes and category nodes, as well as by adjusting a set of attention weights that determine the importance of each stimulus dimension.

**DIVA**. DIVA offers a more generative than discriminative approach to classification learning and deals in distributed rather that localist internal represenations. Learning to classify examples is accomplished by minimizing reconstructive error along the channels of a divergent autoencoder that is comprised of recoding (input → hidden) weights shared for all categories and separate sets of decoding (hidden → output) weights dedicated to each category. Classification judgments are based on which category channel yields the lowest error, i.e., which channel has been tuned to expect (and successfully reconstruct) a set of features like those of the current item.

DIVA is similarity-based in the sense that the model learns, for each category, how to effectively predict feature values for particular regions in recoding space – when an input item projects into a region that is well handled by a category, the reconstructive error in predicting the features will be low. DIVA does not apply Shepard-like stimulus generalization to categorization – an item is likely to belong to a category because its feature values conform to what a category channel has been optimized to successfully recode and decode, not because it is highly similar to a known member of the category.

**Our approach to model comparison**. We compare models based on their ability to account for human generalization after category learning. An important advantage of focusing on generalization performance is that we avoid the traditional reliance on post-hoc fits. In all cases, we first fit DIVA and ALCOVE to averaged learning data from each condition in order to find best-fitting parameters across the full set of conditions. This procedure allows us to separate out the parameter fitting process, so that the generalization performance is genuinely a prediction based on a selected model.

We elected to fit ALCOVE using a grid search over its response mapping ($\phi$), specificity constant (c), association weight learning rate, and attention learning rate parameters. We also fit DIVA using a grid search over the parameters: learning rate, weight range, number of hidden nodes, and a new focusing parameter ($\beta$) that gives DIVA the ability to account for sensitivity to differences in dimension diagnosticity (Kurtz, 2008).

DIVA's focusing parameter ($\beta$) allows it to selectively attend to stimulus dimensions based on the disparity in the output activations for that dimension across category channels. DIVA's focusing mechanism differs significantly from selective attention in ALCOVE in that it does not change the encoding of the stimulus or manipulate the representation learned by the model. DIVA's form of focusing is decisional, rather than perceptual or representational in nature, as it operates at the level of the classification response.

## Experiment 1

This experiment was designed to explore generalization under two conditions: when all stimulus dimensions are diagnostic and equally salient; and when all dimensions are diagnostic, but unequally salient. Figure 1 depicts the two category structures.

Stimulus scaling is an important aspect of our salience manipulation. In order to determine the relation between the stimulus dimensions, we scale the examples in a pairwise similarity study. The similarity study generates a full set of scaled examples, which allows us to manipulate the distance between examples on any dimension. The salience of a dimension can be specified by the distance between the categories on that dimension.

In a pilot study, we explored an extreme case of classification learning in which both stimulus dimensions were diagnostic, but one dimension was much less salient. Participants were generally insensitive to variation in the less salient dimension. In light of these findings, we expected that generalization gradients would show sensitivity given a relatively moderate difference in dimension salience.

**Participants and Materials**. 108 undergraduates from Binghamton University participated in partial fulfillment of a course requirement. Stimuli were rectangles varying in shading and the distance between two lines within the rectangle. Examples were generated at 8 positions on each dimension (8 shading * 8 line spacing = 64 examples). The category structures are depicted in Figure 1 along with sample stimuli.

**Procedure**. Participants were randomly assigned to either the equal salience group or the unequal salience group. In the equal salience condition, the category prototypes were separated by distances of 0.64 and 0.54 on the first and second dimensions (shading and line spacing), respectively. In the unequal salience condition, the category prototypes were separated by a distance of 0.65 and 0.34 on the first and second dimensions. In each condition, there were 4 training examples in each of the two categories.

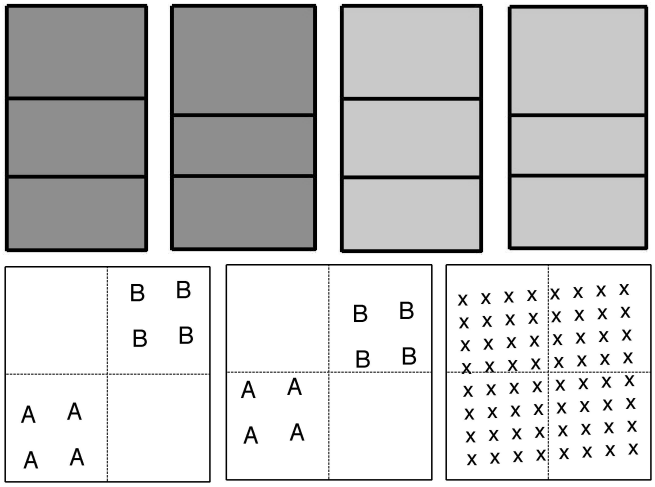

Figure 1: *Top*: Four examples of stimuli (taken from the corners of the stimulus space). *Bottom left & center*: Category structures with equally and unequally salient dimensions. *Bottom right*: Test set used for Experiments 1 and 2. Note that all training items are included in the test set. Positions of examples reflect prior scaling.

Each participant completed 32 learning trials. On each trial, a training item was presented on the computer screen and participants were prompted to make a classification decision by clicking one of two buttons (labeled 'Alpha' and 'Beta'). After responding, participants were given corrective feedback on their response. In the test phase, participants classified the 64 examples sampled across the stimulus space (test set depicted in Figure 1). The 8 training items were also presented during the test phase.

**Gradient Analysis**. In the test phase, participants provide data that yield a generalization gradient of their classification responses. For each participant, we calculated the standard deviation of classification responses at 8 positions on each dimension of the gradient. We then estimated sensitivity to each dimension by calculating the mean of these 8 values. Insensitivity to a dimension is indicated by uniformity of classification responses across that dimension.

**Results and Discussion**. 24 participants were excluded from the subsequent analyses for failing to correctly classify 7 out of 8 training items presented during the test phase. The remaining participants were more than 96% accurate during the last training block in both conditions.

There were significant individual differences in the generalization data. A k-means analysis revealed three profiles based on the sensitivity estimates described above: these were *unidimensional* generalization based on either one or the other stimulus dimension (shading or spacing) and *multidimensional* generalization based on both dimensions. We compared the k-means findings across salience conditions (results are shown in Figure 2).

While a very few participants were sensitive to both dimensions at test, the majority of participants generalized undimensionally. A Fisher's Exact test revealed that the rate of each unidimensional profile differed between salience conditions ($p < .001$). Participants in the unequal salience condition were more likely to be sensitive to the salient dimension (shading) than participants in the equal salience condition.

We observed a bias towards the line spacing dimension in the equal salience group that is not consistent with the scaling. Interestingly, this may reflect a task difference between pairwise similarity and classification learning that renders participants differentially sensitive to our stimulus dimensions.

The main conclusions we can draw from this study of a 'minimal case' category structure are that: 1) participants tended to generalize according to a single dimension despite an optimal diagonal bound; and 2) dimension salience increased the likelihood of the dimension serving as the basis for generalization.

**Modeling Analyses**. We tested DIVA and ALCOVE for their ability to account for these generalization findings. Specifically, we sought to determine whether the models could account for: (1) the tendency of learners to generalize based on a single dimension; (2) the substantial degree of selection of each of the two dimension as the focal one by different sets of learners; and (3) the effect of salience on dimensional sensitivity.

Before generating predictions for generalization, we obtained optimal parameter sets by fitting the models to the aggregate learning data (minimizing the sum of squared deviations, SSD, across learning blocks). We then generated predictions for generalization across a range of optimal

Table 1: Parameter values for ALCOVE and DIVA that best fit all conditions of learning performance in Experiments 1 and 2.

ALCOVE

| | Experiment 1 *SSD < .003* | Experiment 2 *SSD < .06* |
|---|---|---|
| *c (specificity)* | *3.4* | *10.5* |
| *ϕ (response mapping)* | *2.8* | *1.45* |
| *attention learning* | *0.0* | *0.0* |
| *association learning* | *0.1* | *0.3* |

DIVA

| | Experiment 1 *SSD < .004* | Experiment 2 *SSD < .03* |
|---|---|---|
| *number of hidden nodes* | *1* | *1* |
| *β (focusing)* | *20* | *80* |
| *learning rate* | *0.14* | *0.18* |
| *initial weight range* | *+/-0.5* | *+/-1.5* |

parameter sets to gain a full understanding of how the two models performed.

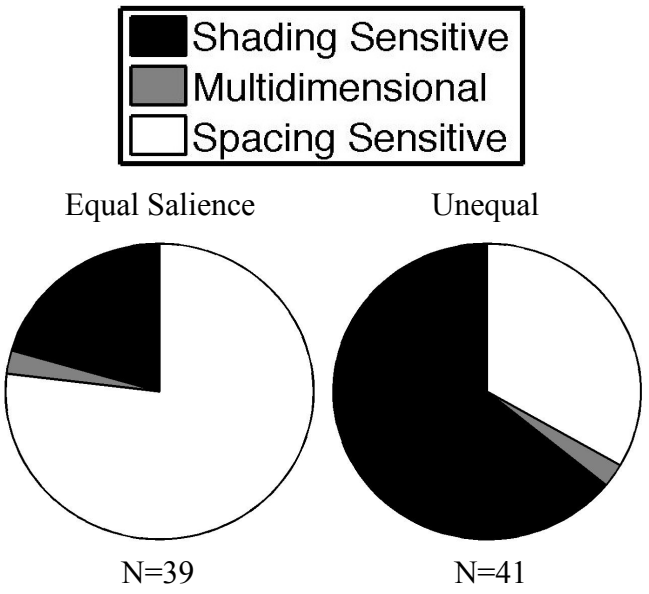


Figure 2: Results of k-means clustering results for Experiment 1. Number of participants shown below each chart.

Both ALCOVE and DIVA provided good fits to the learning data under a range of parameters. The best fitting parameter sets are shown in Table 1. When we tested the models on generalization using these parameters, we found that both models were sensitive to the salience of each dimension, but neither predicted the unexpected bias toward the line spacing dimension that was observed behaviorally.

ALCOVE's attention learning parameter largely governed the model's ability to generalize to a single dimension. ALCOVE produced unidimensional gradients with high levels of attention learning and multidimensional gradients with low levels of attention learning. Given a high attention learning parameter, ALCOVE generalized based on whichever dimension was most salient. We note that ALCOVE lacks any random element such as initial weight values, so the output is deterministic; for this reason, the model does not account for the heavy use of both possible unidimensional rules in the generalization data. Future research will explore generalization using a stochastic version of ALCOVE.

Similar to ALCOVE's use of attention, DIVA's focusing parameter allowed the model to generate either unidimensional or multidimensional gradients. But unlike ALCOVE, DIVA is initialized with random weights on every run. An analysis of results on individual runs revealed that when DIVA's focusing parameter was large and the dimensions were equally salient, the random initial weights sometimes lead to unidimensional generalization based on either dimension. With larger weight ranges, DIVA produced varied distributions of generalization profiles.

Our analysis of DIVA's generalization also revealed that, with a high focusing parameter, the model is more likely to generalize based on a salient dimension than a less salient dimension. This trend resembles the effect of salience that was observed previously. When the dimensions are equally salient, DIVA tends to produce multidimensional profiles at a greater rate than would be predicted given our behavioral findings.

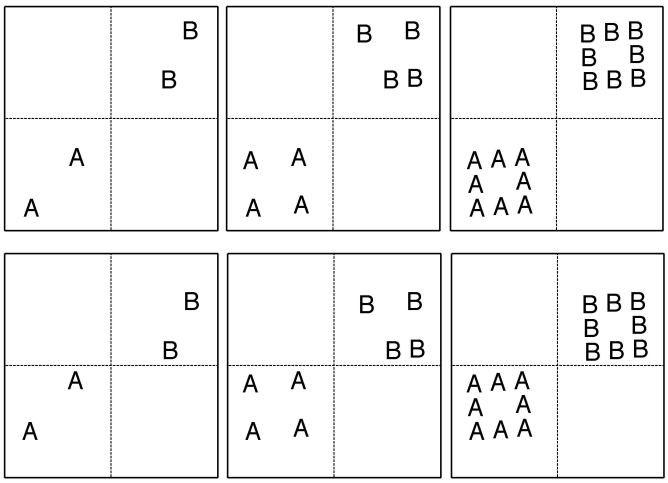

Figure 3: Category structures for Experiment 2.

These modeling results confirm that generalization provides a promising basis for model evaluation. We found that DIVA and ALCOVE produce generalization gradients that are consistent with the salience of each dimension, and that attentional mechanisms allow similarity-based models to generate unidimensional gradients. Furthermore, a random component can partially explain variability in dimensional selection.

## Experiment 2

This study was designed to replicate and extend Experiment 1. As in the first study, we manipulated the salience of dimensions by modifying the distance between the two categories. We extend the design by incorporating category size as a between-participants facto (Figure 3 depicts the category structures that were employed). Category size is a potentially interesting factor in our studies because increasing the number of examples in each category also increases variation in representational demands for exemplar models like ALCOVE without altering the solution that the model is required to find. Furthermore, increases in category size should decrease the memorizability of each example (see Homa, 1984 for background on category size effects).

Our primary predictions were that: (1) generalization after learning would reflect sensitivity to a salient

dimension; and (2) shifts in category size would impact the prevalence of integrated, multidimensional generalization.

Table 2: Distance between opposite-category prototypes on each dimension.

| | Equal Salience | | Unequal Salience | |
|---|---|---|---|---|
| | Shading | Spacing | Shading | Spacing |
| *2 eg* | 0.70 | 0.55 | 0.72 | 0.34 |
| *4 eg* | 0.67 | 0.55 | 0.69 | 0.34 |
| *8 eg* | 0.64 | 0.54 | 0.65 | 0.34 |

**Participants and Materials**. 228 undergraduates from Binghamton University participated in this experiment toward partially fulfillment of a course requirement. The materials were like those used in Experiment 1.

**Procedure**. Participants were randomly assigned to one of six conditions (2 levels of salience x 3 levels of category size). The category structures are depicted in Figure 3. Participants learned a classification based on two, four, or eight unique examples per category.

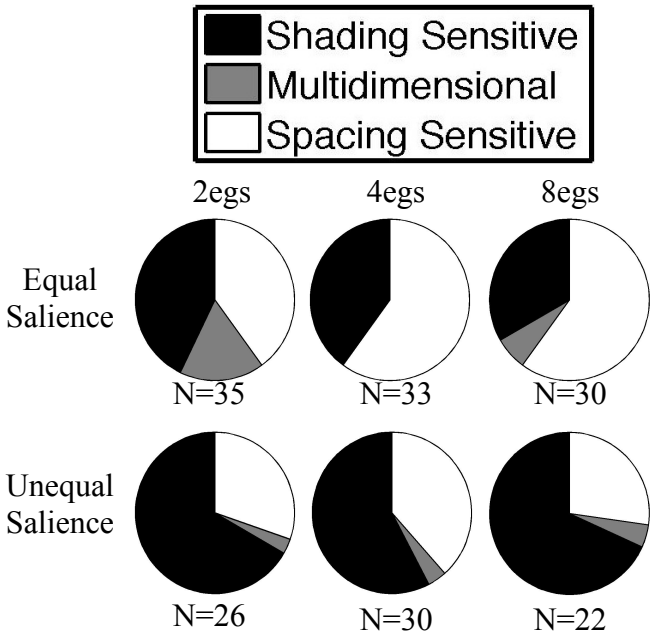


Figure 4: Experiment 2 k-means clustering results. Number of participants shown below each chart.

The salience manipulation was similar to that used in Experiment 1 with one departure – we partially re-arranged the members of the second category so that the category prototypes would more evenly spaced apart in the equal salient condition. The distances between prototypes for each condition are shown in Table 2. All other aspects of the procedure are identical to Experiment 1.

**Results and Discussion**. 56 participants were excluded from subsequent analyses because they made more than one error on training items presented during the test phase. The remaining participants were more than 94% accurate during the last training block.

The analysis of the generalization data was conducted as in the first study. Results are displayed in Figure 4. The data do not reveal an effect of category size on generalization. Consequently, our discussion focuses on the salience manipulation across category size conditions.

As in Experiment 1, the majority of participants generalized to a single dimension. A Fishers Exact test (conducted across size groups) reveals a significant effect of salience ($p < .01$). Participants in the unequal salience group tended to generalize using the salient dimension over the less salient dimension.

We observed the same bias towards the line spacing dimension in the equal salience conditions: our participants were highly sensitive to the line spacing dimension, even when the scaling revealed that the dimensions were equally salient.

**Modeling Analyses**. We again tested DIVA and ALCOVE on their ability to match human generalization performance. In general, the modeling results for Experiment 2 parallel the results of Experiment 1. Both models found good fits to the aggregate learning data, but neither model predicted the unexpected bias towards the line spacing dimension during generalization. Neither model was affected by our category size manipulation. Parameter information can be found in Table 1.

As in Experiment 1, ALCOVE's attention learning parameter allowed it to account for unidimensional generalization. Given a high attention learning parameter, ALCOVE generalized based on whichever dimension is most salient. But due to the lack of a random component, ALCOVE could not account for the use of either single dimension.

As was the case for attention learning in ALCOVE, DIVA's focusing parameter allowed it to account for unidimensional generalization. Replicating our findings from Experiment 1, we found that when DIVA's focusing parameter was large and the dimensions were equally salient, the random initial weights lead to a distribution of generalization profiles based on either or both dimensions. With larger initial weight ranges, DIVA produced more varied patterns of generalization.

The distributions produced by DIVA reflected the salience of the stimulus dimensions. Specifically, DIVA was more likely to generalize using a salient dimension than a less salient dimension. This trend is similar to the effect of salience that we observed behaviorally. Lastly, as in Experiment 1, DIVA tended to produce more multidimensional profiles when the dimensions were equally salient.

## General Discussion

Our behavioral results revealed that: (1) category knowledge tends to be generalized based on a single dimension; and (2) the salience of a dimension affects the probability that it is selected. We compared DIVA and ALCOVE on their ability to account for these findings. We learned that these similarity-based models are sensitive to salience differences between dimensions and can use attention to generate unidimensional gradients. We also found that a random component can help predict arbitrary dimension selection: DIVA's initial weights randomly offset the models salience appraisal and allowed it to generalize to a single dimension.

These results help to establish generalization as an important basis for formal model evaluation. By requiring that models account for generalization and learning based on the same parameter fits, we systematically widen the scope of what models are held accountable for explaining. In our work, generalization proved not only to be area where DIVA and ALCOVE made different predictions, but it also provided an opportunity to reduce our reliance on post-hoc fits by searching for parameters using aggregate learning data. In future work, we plan to conduct simulations using a stochastic modification of ALCOVE in order to determine how well the model matches our distributions of human generalization performance. We also plan to conduct new simulations based on fitting the models to individual learning curves rather than aggregate data.

## Acknowledgments

We would like to thank the members of the Learning and Representation in Cognition (LaRC) Laboratory at Binghamton University.

## References

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

Erickson, M. A. & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, 9(1), 160-168.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), Psychology of learning and motivation (Vol. 18, pp. 49-94). New York: Academic Press.

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

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, 14, 560 –576.

Kurtz, K. J. (2008). Advances in modeling human category learning with DIVA. Presented at the 2008 Annual Meeting of the Psychonomic Society, Chicago, IL.

Levering, K., & Kurtz, K. J. (2010). Generalization in higher-order cognition: Categorization and analogy as bridges to stored knowledge. In M. T. Banich & D. Caccamise (Eds.), Generalization of knowledge: Multidisciplinary perspectives(pp. 175–196). New York, NY: Psychology Press

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

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes* (pp. 149-167). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Nosofsky, R. M. & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375-402.

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

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.