

# Categorizing Fragments of Exemplars: Experimental and Computational Results

Harlan D. Harris (harlan.harris@nyu.edu)

6 Washington Place, 8th Floor  
New York, NY 10003 USA

## Abstract

Analysis of the relative categorization accuracy of whole and partial items may provide deeper insight into the nature of categorization processes. I compared several logical, mathematical and computational models of categorization (including the FLMP, KRES, and variants of the GCM), and ran an experiment to get robust empirical data. The experiment showed a mixture of categorization strategies used by different subjects. When known rule-based strategies were excluded, the remaining subjects responded more accurately than theoretically optimal to single feature stimuli. This result is inconsistent with standard exemplar and connectionist models, and with theoretical models of perceptual integration, but is consistent with an interactive model of categorization that allows top-down feedback to affect representations of partial stimuli.

**Keywords:** concepts and categories; computer simulation; human experimentation; mathematical modeling; neural networks

Decades of work on the basic cognitive processes of categorization has provided sophisticated information about how stimuli are represented, how categories are stored, how similarity is computed, and how categorization choices are made (Ashby & Maddox, 2005). Surprisingly, little work has considered how partial stimuli are categorized, and as far as I know, no work has investigated the mathematical relationships between partial-item and whole-item categorization of stimuli with more than two features.

Should missing properties be ignored for the purposes of categorization? Or, if only 20% of a representation is present, perhaps we should be only 20% confident in a categorization? Most physical objects that have to be categorized during our day-to-day lives are not fully visible, yet are categorized quickly and accurately, as if the missing information were unimportant (Taylor & Ross, 2007). Different categorization models make different predictions about how partial stimuli are treated, so a better understanding of this process would constrain the set of viable theories.

Several studies have tangentially addressed categorization of partial stimuli. Estes, Campbell, Hatsopoulos, and Hurwitz (1989), in a study about base-rate effects, considers categorization of hypothetical diseases given unequally predictive symptoms. Base-rate neglect was found during single-feature tests, with accuracy higher than expected. (Here, I examine the simpler case of non-probabilistic, equally-likely categories, with equally predictive features.) Referencing this study, Estes (1994) notes that partial stimuli can be encoded in a model using either a *union rule* or an *intersection rule*. In the former case, missing dimensions are treated as a mismatch in the similarity computation; in the latter case, missing dimensions are treated as a match.

Two recent papers (Verguts, Ameel, & Storms, 2004; Taylor & Ross, 2007) have considered the effects of partial stim-

uli on category *learning* (not *categorization*). Taylor and Ross (2007) found that learning with partial stimuli yields more extensive learning about category structure. Verguts et al. (2004), in a paper focusing on problems with geometric theories of similarity and the effects of redundant features, also considered learning with partial stimuli. They described a version of the ALCOVE model (Kruschke, 1992) with an additional parameter that allows it to fit their data, but do not analyze this issue extensively.

The work reported here was spurred by research considering how changes in the number of dimensions affects learning of family resemblance categories (e.g., Table 1, left). Hoffman and Murphy (2006) and Hoffman, Harris, and Murphy (2008) considered learning of such categories and found that more dimensions were learned when more dimensions were present, under a variety of conditions. As part of their analysis, Hoffman et al. (2008) tested several computational models of category learning. A standard independent-feature recognition model, the Fuzzy Logical Model of Perception (FLMP, Oden & Massaro, 1978), predicts a precise relationship between Whole Item (WI) and Single Feature (SF) response accuracy, with SF accuracy lower than was empirically found. Other straightforward models had similar problems. However, the Knowledge Resonance (KRES, Rehder & Murphy, 2003) model was able to account for the relatively high SF accuracy level, as well as other details of the experimental results.

The results of Hoffman et al. (2008) raise as many questions as they answer, so this work moves beyond some of that work's limitations. First, I have extended the analysis to consider exemplar models and Estes' union and intersection rules. Second, a new experiment expands testing to allow analysis of individual subject strategies. Third, early test phases allow investigation of responses throughout learning.

## Models and Simulations

The task given to the models (and the participants, below) was to distinguish two categories of stimuli. The abstract category structure is shown in Table 1 (left). Each training stimulus was made up of four features that are predominant in one category and one feature that is predominant in the other category. All dimensions are equally predictive of the category. There were three types of test items (Table 1, right): whole items (D1-D5, K1-K5), prototypes (D0, K0), and single-features (DF1-DF5, KF1-KF5).

## Rule-based Models

Learning can induce logical rules that describe the categories. Different rules make different predictions about test item ac-

Table 1: Abstract Category Structure

Training		Testing	
Item	Features	Item	Features
D1	10000	training items plus:	
D2	01000	D0	00000
D3	00100	K0	11111
D4	00010	DF1	0---
D5	00001	DF2	-0---
K1	01111	DF3	--0-
K2	10111	DF4	---0-
K3	11011	DF5	----0
K4	11101	KF1	1---
K5	11110	KF2	etc.

Note. “-” symbol indicates no feature presented for the given dimension. In the experiment, items D0 and K0 were tested twice as often as the other items.

curacies. First, using a 1-D rule, four of the five dimensions are ignored while one dimension is used as the cue to categorization. Prototype accuracy would then be 100%, WI accuracy would be 80% (reflecting the cue validity of each dimension), and SF accuracy would be 60% (100% accuracy on the attended-to dimension, averaged with 50% accuracy on the other dimensions.)

The simplest rule that always correctly categorizes whole items is to attend to three of the five dimensions and use a 2-of-3 rule. For example, if the stimulus is 11011, and the first three dimensions are those attended, then there are more 1s than 0s, and the appropriate response would be made. Under this rule, prototype and WI accuracy would be 100%, while SF accuracy would be 80% (100% on three dimensions, and 50% on two).

### FLMP/Bayes/Network Models

Next, consider the FLMP, a commonly-used and successful model of how independent sources of evidence may be combined in recognition (Massaro, 1987; Massaro & Friedman, 1990). The FLMP assumes that stimulus features are represented as the degree to which they support each of the categories, on a continuous scale from 0 (completely false), to .5 (no evidence), to 1.0 (completely true). Category response probabilities are then calculated using a standard choice rule:

$$P(A_k|X = X_i, Y = Y_j) = \frac{x_k y_k}{\sum_{i=1}^m x_i y_i}. \quad (1)$$

The FLMP is a Bayesian rule for combining evidence, if the representation of the features is precisely equal to the cue validity (Cohen & Massaro, 1992).

A mathematical property of the FLMP is that relative accuracy of responses to whole stimuli can be predicted from the accuracy of responses to partial stimuli, or vice-versa. When each dimension is equally predictive, as in the current category structure (Table 1), this is particularly easy to write:

$$W = \frac{f^n}{f^n + (1-f)^n} = \frac{(f/(1-f))^n}{(f/(1-f))^n + 1}, \quad (2)$$

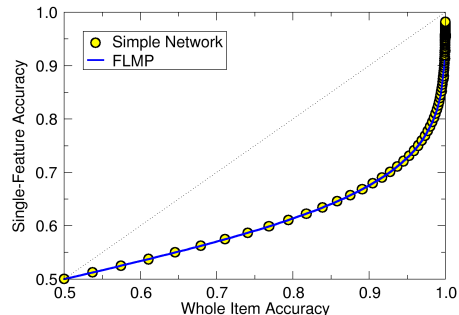


Figure 1: Relative accuracy of simple 1-layer network models on the whole-item and single-feature tests, over a range of parameters. Solid line is the predictions of the FLMP.

$W$  is the probability of responding correctly to a whole item,  $f$  is the probability of responding correctly to any single feature, and  $n$  is the number of dimensions that make up a whole item. The inverse of this function predicts SF accuracy from WI accuracy. (Note that here, as the whole items such as 11110 have four supporting features and one contradictory feature,  $n = 3$  instead of  $n = 5$ .)

A further equivalence is between the FLMP and a standard one-layer connectionist network (Cohen & Massaro, 1992; Massaro & Friedman, 1990). This type of network has one output node per potential category response, where each node computes the weighted sum of the stimulus representation, then “squashes” the result through a sigmoidal function to get a response probability. The weights of the network can be changed according to a supervised learning rule. When only two category responses are possible, this sort of network will make exactly the same predictions as the FLMP, for any stage of learning and any (non-degenerate) set of weights.

I implemented a one-layer network model with fixed weights representing category prototypes, as if the model had been trained to perfection, and tested it on whole items and single features. Five input nodes were connected to a single output node with all weights  $w_i = 1$ . Whole item stimuli were of the form  $I = [1, 1, 1, 1, -1]$ , while SF stimuli were of the form  $I = [1, 0, 0, 0, 0]$ . The output node computed the function  $O = (1 + \tanh(\beta \cdot I \cdot w))/2$ , where  $\beta$  is a parameter indicating the sharpness of the sigmoid, and  $O$  was the probability of responding consistently with the 11111 prototype. Note that imperfect learning would be represented in the model by  $w < 1$ , which is mathematically equivalent in this case to scaling  $\beta$ , so varying  $\beta$  simulates testing the model with varying amounts of training. Figure 1 shows the predictions of this simple network model, with  $\beta$  ranging from 0 to 2, along with the curve defined by Equation 2. Clearly, the model’s simulated predictions exactly line up with the FLMP’s theoretical predictions. Single feature accuracy is significantly lower than the accuracy on the whole items, throughout the parameter space of the model.

## GCM

The Generalized Context Model (Nosofsky, 1986) is an exemplar model of categorization, and so assumes that categorizations are made by comparing the stimulus to stored exemplars, with responses based on a function of those similarity computations (Medin & Schaffer, 1978). The equations that describe the GCM are as follows:

$$d_{ij} = \left[ \sum_{k=1}^m (w_k |x_{ik} - x_{jk}|)^r \right]^{1/r}, \quad (3)$$

$$s_{ij} = e^{-\lambda d_{ij}}, \quad (4)$$

$$s_{ic} = \sum_{j \in c} s_{ij}, \quad (5)$$

$$\theta_{ic} = \frac{s_{ic}^\gamma}{\sum_d s_{id}^\gamma}. \quad (6)$$

The first equation describes the psychological distance between stimuli  $i$  and  $j$ , the second describes the similarity between stimuli  $i$  and  $j$ , the third describes the similarity between stimulus  $i$  and category  $c$  (containing several exemplars), and the fourth describes the decision rule. The  $r$  parameter varies depending on the representation of the stimuli, and is typically set to  $r = 1$  for separable dimensions, such as are used here, or  $r = 2$  for integral dimensions (but see below). As all dimensions are equally salient, all weights  $w_k = 1/n$ .  $\lambda$  (often called  $c$ ) is a parameter that determines the sharpness of the generalization or typicality gradient.  $\gamma$  is a response scaling parameter that related category membership probabilities and response proportions; when  $\gamma = 1$ , the model probability-matches (Ashby & Maddox, 1993).

As did Verguts et al. (2004), I added a parameter  $m$  to the GCM that allows missing values to be dealt with during similarity comparisons. Assuming the values of dimensions are normally  $-1$  or  $+1$ , missing dimensions from the stimulus are replaced by  $m$  times the value from the exemplar. If  $m = -1$ , missing dimensions are treated as maximally different from the point of view of the similarity calculation, ala Estes' Union rule. If  $m = +1$ , missing dimensions are ignored from the point of view of the similarity calculation, ala Estes' Intersection rule. Intermediate values have intermediate effects on similarity calculations.

Figure 2 shows the results of simulating the GCM (+  $m$ ) on the test stimuli in Table 1, with the training stimuli as stored exemplars. The  $\lambda$  parameter was varied over the range  $[0, 10]$ , while the  $\gamma$  and  $r$  parameters were set at either 1, their usual values for this task, or 2. When  $r = 1$ , as is typically used when stimuli have separable dimensions, the GCM predicts SF accuracy slightly below the FLMP predictions. When  $r = 2$ , however, SF accuracy can be higher than FLMP predictions. Curves with  $\gamma = 2$  are similar.

## KRES

The KRES model (Rehder & Murphy, 2003) is an interactive activation, constraint satisfaction model. Features and

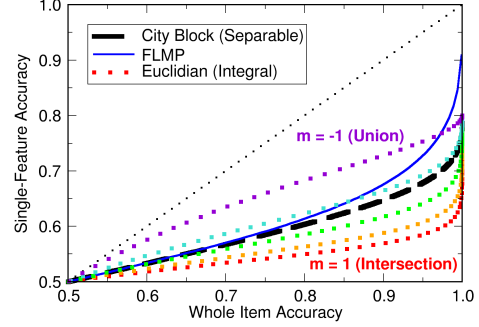


Figure 2: Relative accuracy of GCM models on the whole-item and single-feature tests, with  $\gamma = 1$  and varying  $\lambda$  parameter. Using  $r = 1$  (city block) similarity metric,  $m$  had no effect (black line). Using  $r = 2$  (Euclidian) similarity metric, effect of  $m$  is shown by different colored lines.

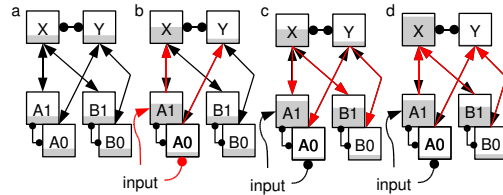


Figure 3: Process of top-down feedback on category activation in KRES. From (a) initial weak universal activation, (b) input affects bottom-up activation of input (A1 and A0) and category (X, Y) nodes, (c) feeds back from category node X to both presented and non-presented input nodes, then (d) reverberates back up to increase category activation further.

categories that have associations are linked together by bidirectional weighted connections. Activation moves along the links in top-down, bottom-up, and lateral directions, eventually converging to a pattern that represents the most consistent interpretation of the stimulus, given everything else the model has represented. (KRES features not used here include representations of prior knowledge, learnable connections between nodes, and the use of exemplar nodes, Harris & Rehder, 2006.) KRES has been shown to account for a variety of effects in categorization, category learning, and the interaction of learning with prior knowledge (Rehder & Murphy, 2003; Harris & Rehder, 2006; Hoffman et al., 2008).

The mathematics that describe the implementation of KRES are presented in Rehder and Murphy (2003), and are similar to other interactive activation models (McClelland & Rumelhart, 1981). A key property of KRES is that top-down feedback can cause missing dimensions to be “filled in,” as if a feature were actually present in the stimulus. Figure 3 illustrates the mechanism in a two-dimensional case. Stimulus “1-” would be treated almost as if it were “11”.

When the KRES model simulates the task (Table 1), SF response accuracies tend to be *above* the FLMP ratio (Fig-

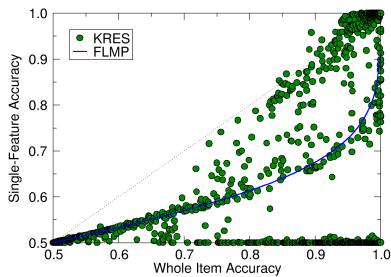


Figure 4: Relative accuracy of 5000 KRES models on the whole-item and single-feature tests, with randomly varying weight and activation parameters.

ure 4). The model had five pairs of input units, connected to category prototype nodes with equally-weighted connections. The magnitude of fixed excitatory and inhibitory weights, as well as a parameter describing the sharpness of node responses, were varied randomly across a wide range. Each stimulus was provided to the input layer in turn (for the two feature nodes representing each dimension, 1 was encoded as  $+1$ ,  $-1$ , 0 was encoded as  $-1$ ,  $+1$ , and  $-$  was encoded as  $0$ ,  $0$ ). Once the network settled, the category activations were converted into response probabilities.

Three patterns are evident (Figure 4). In one pattern, accuracy on whole items is relatively good yet SF accuracy is at chance. Analysis showed that model parameters were large in magnitude and the network was in an unstable part of its parameter space. In a second pattern, KRES performance is similar to that of the simple network, suggesting that some parameter settings yield little top-down feedback. (With only bottom-up activation flow, KRES’s computations are nearly the same as those of the simple network.) In the third pattern, KRES performance is significantly above the FLMP curve once whole-item performance exceeds about 70%. Top-down feedback enhances activation of both presented and nonpresented features, using them to boost response accuracy above what would be expected from an independent feature model.

These results show that the result observed by Hoffman et al. (2008), a relatively high SF/WI ratio, could potentially be accounted for by either the KRES model or by the GCM (but only if the stimuli are integral). To evaluate these predictions, I now turn to empirical methods.

### Experiment: Whole Items and Single Features

This experiment was designed to replicate and extend the Hoffman et al. (2008) result, and to test the computational models discussed above. As in the earlier experiment, subjects learned about stimuli with a family resemblance structure (Table 1) and were then tested.

The major factor of interest was a within-subjects contrast—the relationship between WI and SF accuracy—but I varied two factors between subjects to get a broader set of data. 1/3

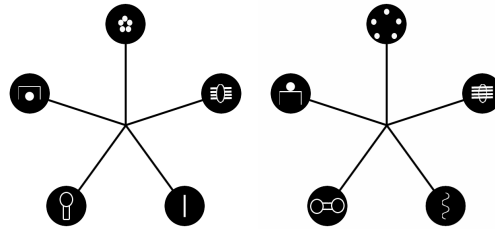


Figure 5: Prototypes of the Dax and Kez categories used in the Experiment.

of subjects were tested following blocks 2 and 8, 1/3 were tested following blocks 5 and 8, and the remaining 1/3 were tested only following block 8. This allowed an examination of performance earlier in learning, when overall accuracy was expected to be lower. Second, half the subjects did “simple” test blocks, categorizing each of the 24 samples twice, while half of the subjects did “complex” test blocks, categorizing each of the 24 samples only once, but also making a frequency estimation that was expected to provide a more continuous dependent measure.

### Method

**Participants and Design** 49 NYU undergraduates participated in the experiment for course credit. Of these, the data from 10 participants were not included due to equipment failure (2), failure to reach a performance criterion (7), or failure to follow instructions (1). Participants were randomly assigned to the two between-subjects factors, *test blocks* (2+8, 5+8, or 8), and *test type* (simple, complex).

**Materials** Stimuli were star-like “ceremonial symbols,” as used previously by Hoffman (2007) (see Figure 5). Participants learned to distinguish the symbols from the “Dax” tribe from symbols from the “Kez” tribe. The abstract category structure is shown in Table 1 (left). All dimensions were equally predictive, and pretesting by Hoffman (2007) indicated that each dimension was about equally salient.

During test phases, subjects responded to the training items, plus additional test items (Table 1, right). Single-feature stimuli were presented with only a single line of the star visible. Prototype stimuli were presented twice as often as other stimuli.

**Procedure** Each participant performed eight blocks of training and either one or two blocks of testing. Each training block consisted of a single training trial for each of the 10 stimuli. Responses had no deadline, and feedback was provided for 4 s.

In the test blocks for the simple test condition, each test item was presented in random order twice, while for the complex test condition, each item was presented once. In the complex condition, following the categorization response, participants provided a frequency estimate: if they had found 100

Table 2: Average Test Accuracy and Frequency Estimates

	Accuracy			Frequency		
	Proto.	WI	SF	Proto.	WI	SF
Test 2	.93	.75	.72	67	72	60
Test 5	.93	.79	.72	72	72	54
Test 8	.98	.91	.84	80	75	69
2+8	1.00	.92	.86	85	83	81
5+8	.97	.91	.82	83	76	63
8	.96	.90	.84	68	64	63

Note. Proto. = Prototype; WI = Whole Item; SF = Single Feature. Frequency estimates are out of 100 items of the specified type. Last three rows show Test 8 accuracy for each *test blocks* condition.

identical symbols, how many of them would be from the culture of their classification response?<sup>1</sup>

## Results

The two major concerns regarding the learning phase of the experiment are whether subjects learned to a reasonable level of proficiently, and whether the presence of the early test phases substantially impacted what was learned. Following training, accuracy had increased from chance to .897, and a mixed-design ANOVA<sup>2</sup> with blocks, test type, and test blocks as factors, showed the expected main effect of blocks,  $F(7, 231) = 30.88 > 2.05, \eta_p^2 = .48$ . As for the effect of the early test phases, the learning curves had somewhat different shapes depending on when subjects performed tests. Training accuracy increased somewhat more sharply following early test blocks, significantly so following the block 5 test,  $F(1, 38) = 4.28 > 4.10, \eta_p^2 = .10$ . However, by the final block of training, the accuracy differences had entirely disappeared. Additionally, the response patterns on test 8 were not affected by the presence or absence of earlier tests,  $F_s(1, 36) < 1$ , as shown in Table 2 (left). These questions answered, we may now proceed with the analysis of the key data from the test blocks themselves.

Figure 6 shows the relationship between WI and SF accuracy during testing. Several clusters of points can be seen, representing participants using different strategies.

First, with  $WI = .8$  and  $SF = .6$ , are subjects who may be using a single dimensional rule to categorize. (See the Rule-based Models section, above.) Four learners had this accuracy patterns in test 8. Second, with  $WI = 1.0$  and  $SF = .8$ , are subjects who may be focusing on three dimensions and ignoring the other two. The 2-out-of-3 rule discussed above is an optimal strategy for learning the whole items of this task.

<sup>1</sup>A more traditional typicality rating would not be useful here, as broken symbols were not typical of the training items, and the quantitative relationship between single-feature estimated frequency and whole-item estimated frequency is the measure in question.

<sup>2</sup>Statistics are reported by comparing the computed test statistic with the threshold for significance, assuming  $p = .05$ , and reporting the partial eta-squared measure of effect size.

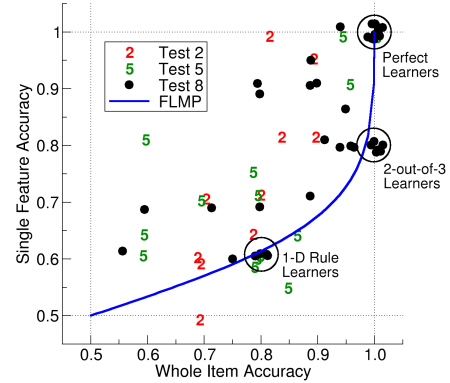


Figure 6: Relationship between WI and SF accuracy on test phases. Data points are randomly jittered to reduce overlap. Blue line indicates theoretical FLMP performance. Circled areas indicate performance predicted by specific rules.

However, on SF tests, subjects using this strategy should have only 80% accuracy. Five learners had this accuracy pattern in test 8. Finally, some learners had perfect accuracy on both WI and SF tests. Any number of strategies would yield this result, and there is no way of distinguishing between them. 12 learners had perfect accuracy by the end of training.

The remaining 38 points on the graph represent test accuracies of participants who (I conservatively assume) were not using the 1-D or 2-of-3 strategies, and who did not have perfect performance. The mean of this subset of the data was  $WI = .789, SF = .750$ . 35 tests were above the FLMP line, and only 3 were below the line. This was statistically significant by a Wilcoxon sign test,  $\chi^2(1) = 26.95 > 3.84$ .

There are thus four patterns of test results. Participants can use 1-D rules or 2-of-3 rules, they can respond perfectly, or they can respond imperfectly but with SF accuracy higher than predicted by independent feature models like the FLMP.

The final analysis is to consider the frequency estimates from the complex test phases. Mean frequency estimates are shown in Table 2 (right). If subjects are frequency matching, such that their response proportions are equal to their frequency estimates, then frequency should parallel accuracy. Although the majority of sensical responses (16 of 23) were above the (frequency matching) FLMP predictions, examination of the data shows that many participants did not make frequency estimates correctly. For example, several subjects gave frequency estimates below 50% (in two cases as low as 20%), despite the fact that there were only two categories, and a response below 50% would imply the opposite category response from what they actual gave. Therefore, the frequency estimation data should be treated with caution.

## Discussion

Different models predict different patterns of SF/WI accuracy ratios. The simple network, FLMP, and related models all predict a strict Bayesian relationship, with SF accuracy rel-

atively low until WI accuracy is very high. Under some circumstances, when stimulus dimensions are integral, the GCM can predict higher SF accuracy. The KRES model predicts a high SF accuracy by virtue of its top-down feedback mechanism. An empirical test of these models found that individual subjects seemed to choose a variety of strategies, including simple decision rules, but that a substantial proportion did not use an identifiable strategy. These subjects almost always had a SF/WI ratio higher than predicted by the FLMP class of models. As the stimuli used in the experiment did not involve integral dimensions, but instead used very separable dimensions, it seems unlikely that the GCM can account for the data. Only the KRES model and its top-down feedback has the potential to account for the results.

This work has two major consequences. First, it continues a line of research that has shown that categorization may be more a process of constraint satisfaction, influenced by top-down, causal, and knowledge-related factors, than a simple similarity comparison (Harris & Rehder, 2006; Murphy & Medin, 1985; Rehder, 2003; Rehder & Murphy, 2003). Second, it supports a novel new line of research that considers partial stimuli as an important test of categorization theories (Hoffman et al., 2008; Taylor & Ross, 2007; Verguts et al., 2004). Critically, the work reported here (and (Hoffman et al., 2008)) is the first to mathematically examine the relationship between whole and partial item categorization accuracies. Future experimental and theoretical work will further examine this issue, with a goal of identifying further constraints on models of categorization.

### Acknowledgments

This work was supported by NIH grants F32MH076452 and MH041704. Thanks to Bob Rehder, Gregory Murphy, and Aaron Hoffman for helpful suggestions, and for Danielle Blinkoff and Renee Fultz for assistance running subjects.

### References

- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372–400.
- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.
- Cohen, M. M., & Massaro, D. W. (1992). On the similarity of categorization models. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 395–448). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 556–571.
- Harris, H. D., & Rehder, B. (2006). Modeling category learning with exemplars and prior knowledge. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1440–1445). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hoffman, A. B. (2007). *Attention dynamics in category learning*. (Unpublished doctoral dissertation, New York University.)
- Hoffman, A. B., Harris, H. D., & Murphy, G. L. (2008). Prior knowledge enhances the category dimensionality effect. *Memory & Cognition, 36*, 256–270.
- Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 301–315.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.
- Massaro, D. W. (1987). Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception*. Cambridge, England: Cambridge University Press.
- Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*, 225–252.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Pt. 1. An account of basic findings. *Psychological Review, 88*, 375–407.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology, 115*, 39–57.
- Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172–191.
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science, 27*, 709–748.
- Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10*, 759–784.
- Taylor, E. G., & Ross, B. H. (2007). *Classifying partial exemplars: Seeing less and learning more*. (Manuscript in preparation)
- Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of cognition. *Memory & Cognition, 32*, 379–389.