Running Head: MECHANISTIC AND RATIONAL APPROACHES

Putting the Psychology Back into Psychological Models: Mechanistic vs. Rational Approaches

Yasuaki Sakamoto

Stevens Institute of Technology

Matt Jones

The University of Texas at Austin

Bradley C. Love

The University of Texas at Austin

Abstract

Two basic approaches to explaining the nature of the mind are the rational and mechanistic approaches. Rational analyses attempt to characterize the environment and the behavioral outcomes that humans seek to optimize, whereas mechanistic models attempt to simulate human behavior using processes and representations analogous to those used by humans. We compared these approaches on their accounts of how humans learn the variability of categories. The mechanistic model departs in subtle ways from rational principles. In particular, the mechanistic model incrementally updates its estimates of category means and variances through error-driven learning, based on discrepancies between new category members and the current representation of each category. The model yields a prediction, which we verify, regarding the effects of order manipulations that the rational approach does not anticipate. Although both rational and mechanistic models can successfully postdict known findings, we suggest that psychological advances are primarily driven by consideration of process and representation, and that rational accounts trail these breakthroughs.

Putting the Psychology Back into Psychological Models: Mechanistic vs. Rational Approaches

Two basic approaches to explaining the nature of the mind are the rational and mechanistic approaches. Rational analyses attempt to characterize the environment and the behavioral outcomes that humans seek to optimize. The rational approach holds that people are adaptive and learn (at the individual or species level) to behave optimally given the nature of the environment (i.e., given available information or statistics). The formal product of a rational analysis is an abstract mathematical model (often Bayesian) that details the behavioral strategies that optimize some cost function given the environment. Such models do not make recourse to how people actually process and represent information, but are instead abstract.

Considerations of the environment and optimality also resonate with adherents of the mechanistic program, but unlike a rational model the main goal of a mechanistic model is to simulate human behavior using the same mechanisms (i.e., analogous processes and representations) as those that support human behavior. The mechanistic program seeks to reverse engineer the human brain and peer inside the black box. The issues of primary importance to the mechanistic program are how people represent and process information.

One common criticism of mechanistic approaches is that they lead to ad hoc explanations that lack the elegance and clarity of models derived from rational analysis. To the extent that two models converge on a common set of predictions, the more transparent and mathematically motivated model should be favored. However, in this paper, we argue by way of demonstration that mechanistic and rational models are likely to diverge in important ways once the full entailments of the mechanistic model are appreciated. In other words, mechanistic models can motivate predictions beyond those of a successful rational analysis. Thus, mechanistic models are properly understood as driving theory advancement, rather than as bastardized instantiations of more abstract rational analyses. Of course, both rational and mechanistic accounts can be constructed after the fact to account for any dataset. Our key meta-scientific argument is that mechanistic models are best suited for deriving surprising behavioral predictions.

To offer tentative support for our meta-scientific argument and to make an independent empirical contribution, we conduct two studies examining how people learn about the variance of categories. The domain we chose is a simple category learning task in which mechanistic and rational accounts are already fleshed out. An initial study exploring the role of category variability in generalization suggests obvious models from within both perspectives. The two models are largely in accord, but an examination of how the mechanistic model builds internal representations of the categories in response to corrective feedback suggests a second experiment in which the predictions of the two accounts diverge and for which the results support the mechanistic account. Empirically, we demonstrate how people's impressions of category variability can be strongly affected by manipulating the order in which category members are experienced.

The unique predictions of the mechanistic model follow from the insight that people incrementally build representations in memory rather than from any insight into the structure of the environment. In effect, the mechanistic model suggests revision of the rational account—a direction of theory development opposite that advocated by proponents of rational analysis.

Experiment 1

Fried and Holyoak (1984) found that after training on two contrasting categories of unequal variance subjects tend to classify intermediate items into the higher-variance category. This sensitivity to category variance has been verified in subsequent learning studies (e.g., Cohen, Nosofsky, & Zaki, 2001; Hahn, Bailey, & Elvin, 2005). Preferences in generalizing to high-dispersion categories have also been found in experiments that tap pre-existing knowledge and categories (Rips, 1989), as opposed to utilizing learning procedures.

Experiment 1 refines aspects of previous learning studies. In Experiment 1, subjects learned to classify lines varying in length into one of two categories. The design is illustrated in Figure 1. Learning items are illustrated as dark triangles. The six items (L1–L6) forming one category are less variable than the six items (H1–H6) forming the contrasting category. Following learning, subjects classified a variety of items, including some items that were not experienced during learning, such as item N6. These novel items are tests of how subjects generalize. Item N6 is of particular interest as it is midway between the nearest trained members (L6 and H1) of the low- and high-dispersion categories.

To foreshadow, our results replicate previous findings indicating that people generalize border items to the high-dispersion category. After the methods and results are presented, mechanistic and rational models are derived and fit to the data.

Methods

Fifty University of Texas undergraduates learned to correctly assign 12 line stimuli (represented by dark triangles labeled L1–L6 and H1–H6 in Figure 1) into category A or B through trial-by-trial classification learning with corrective feedback. The members of one category (L1–L6) varied relatively little in their lengths, whereas the members of the other category (H1–H6) were highly variable. The stimulus lengths in pixels (100 pixels = 33.25 mm) are presented in Figure 1. To eliminate possible influences of absolute line length on performance (Ono, 1967), whether the high-dispersion category had longer or shorter lines than the low-dispersion category was counterbalanced between subjects (see Figure 1). The border item N6 has the same length in both conditions.

On each training trial, one line was presented horizontally at the center of a display and the text “Category A or B?” appeared at the top left corner of the display. After responding A or B, subjects received visual (e.g., “Right! The correct answer is A.”, “Wrong! The correct answer is B.”) and auditory corrective feedback (a low-pitch tone for errors and a high-pitch tone for correct responses). The visual feedback (presented at the bottom left corner of the display) and the stimulus were displayed for 2000 ms after subjects responded. Subjects completed 10 blocks of training trials. A block comprised presentation of every training item in a random order. The density curves shown in Figure 1 are illustrative of possible mental representations, as discussed below, and do not indicate information about the frequency of presentation during the experiment.

Following training, subjects answered three addition problems to prevent rehearsal of information from the learning phase. Finally, subjects completed two blocks of transfer classification. In each transfer block, subjects classified the 12 studied and 11 novel items (represented by light triangles labeled N1–N11 in Figure 1) in a random order as they did in the training phase except no corrective feedback was provided in the transfer phase. Our main interest was subjects' performance on the border transfer item (N6) that was midway between the nearest studied members (L6 and H1) of the two categories.

Results

Border item N6 was more likely to be classified into the high- than into the low-dispersion category. As shown in Figure 2, averaged across the two transfer blocks, subjects assigned the border item to the high-dispersion category with greater-than-chance probability (.69 vs. .5), $t(49) = 3.86$, $p < .001$. In the first transfer block, more subjects (33 of 50) classified the border item to the high-dispersion category than was expected by chance, exact binomial $p = .033$ (two-tailed). The same pattern (36 of 50) was found for item N6 in the second transfer block, exact binomial $p = .0026$ (two-tailed).

Rational and Mechanistic Models

Straightforward rational analyses, whether following Maximum Likelihood (e.g., Fried & Holyoak, 1984) or Bayesian (e.g., Tenenbaum & Griffiths, 2001) canon, converge in their account of Experiment 1. A rational analysis of Experiment 1 suggests a model that estimates the true mean and variance of each category based on the unbiased integration of information conveyed by the training items. Although these estimates can be made incrementally (e.g., the current trial's posterior distribution serves as the next trial's prior distribution in a Bayesian scheme), they are equivalent to estimating the mean and variance based on all experienced items (i.e., perfect and unbiased memory). From these estimated means and variances, the probability that a novel item belongs to each category can be calculated and the item assigned to the more likely category. One such model is the unequal variance signal detection model (Green & Swets, 1966; Maddox & Ashby, 1998) when the standard deviation and mean of each category distribution are estimated from all previous learning trials. These rational models correctly predict that the border item N6 will be assigned to the high-dispersion category.

Experiment 1's results are problematic for existing mechanistic models. Two of the most successful classes of mechanistic models of category learning are prototype models and exemplar models. Prototype models represent each category by its prototypical (or average) member (Reed, 1972). Exemplar models represent categories by storing all encountered examples (Medin & Schaffer, 1978). Both models classify new instances based on their relative similarity to these stored category representations. Both prototype and exemplar models strongly predict that subjects will classify border item N6 into the low-dispersion category because the same similarity metric is used for the low- and high-dispersion categories, and the prototype for the low-dispersion category is closer to N6 as are the exemplars forming the low-dispersion category.

When a mechanistic model fails, it is revised by considering how the mechanism fails (in this case the similarity metric must differ for the two categories) and what aspects of other successful models can be adopted. One straightforward approach is to adapt existing similarity-based models that utilize error-driven learning (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004) so that they can develop different generalization gradients for each category (cf. Nosofsky & Johansen, 2000; Sakamoto, Matsuka, & Love, 2004). Tightening a generalization gradient leads to similarity falling off faster with increasing differences (i.e., distance) between stimuli.

To facilitate comparison, the revised model will principally differ from the aforementioned rational models in that the mechanistic model does not have perfect memory for the training items. Instead, it sequentially updates its representation of each category (both mean and dispersion) based on the current stimulus using error-driven learning. Like the rational models, the mechanistic model represents each category in terms of its mean and variance. This information is represented by a cluster for that category (cf. Anderson, 1991). The cluster tracks the prototype of the category while also encoding its variability. Activation of cluster *i*, a_i, represents the strength of evidence that a stimulus belongs to category *i*, and is a Gaussian function of the presented stimulus value, *x*:

$$a_i = \frac{1}{\sqrt{2\pi}s_i} e^{-\frac{(x-m_i)^2}{2s_i^2}} \tag{1}$$

where m_i and s_i are the cluster's mean and standard deviation, respectively. The generalization gradient of a category is captured by s_i. The response probability for each category is proportional to the activation of the corresponding cluster (i.e., the probability matching response rule).

Cluster means and standard deviations are learned by gradient descent on an error, $E = ½(t_i - a_i)^2$:

$$\Delta m_i = -\varepsilon_m \frac{\partial E}{\partial m_i} = \varepsilon_m (t_i - a_i) \frac{x - m_i}{s_i^3 \sqrt{2\pi}} e^{-\frac{(x-m_i)^2}{2s_i^2}} \qquad (2)$$

$$\Delta s_i = -\varepsilon_s \frac{\partial E}{\partial s_i} = \varepsilon_s (t_i - a_i) \frac{(x - m_i)^2 - s_i^2}{s_i^4 \sqrt{2\pi}} e^{-\frac{(x-m_i)^2}{2s_i^2}} \qquad (3)$$

where ε_m and ε_s are learning rates for cluster means and standard deviations, respectively, and t_i is the feedback to cluster i, equal to α if the stimulus is in category i and 0 otherwise. Cluster means are initialized at the value of the first presented stimulus in each category, and standard deviations are initialized at s_0.

The mechanistic model was trained and tested in a trial-by-trial fashion paralleling the procedure used with the human subjects. Figure 3 illustrates the dynamics of the model simulated on Experiment 1. This figure is based on an average over 10,000 separate runs, using the parameter values $s_0 = 20$, $\alpha = .05$, $\varepsilon_s = 70000$, and $\varepsilon_m = 98000$. These parameters were chosen to fit data from Experiments 1 and 2 simultaneously, but the qualitative results of both experiments were robust to the majority of the parameter space explored.

Prior to training, both clusters have the same standard deviation of 20 and border item N6 is closer to the cluster representing the low-variability category. Thus item N6 should initially be assigned to the low-variability category as it more strongly activates that category's cluster. To confirm this intuition, twenty-five University of Texas undergraduates were shown the two category prototypes (no other training) and chose the category to which the border stimulus belonged. In this single triad task, 22 of 25 subjects preferred to classify border item N6 into the low-dispersion category (i.e., the nearer prototype), exact binomial p = .00016 (two-tailed). Clearly, Experiment 1's categorization training strongly reversed people's initial preferences.

Cluster dispersions are adjusted to maximize within-category activation and to minimize unwanted activation from items belonging to the opposing category. These dynamics lead to learned standard deviations of 9.7 for the low-dispersion category and 22.7 for the high-dispersion category (averaged across simulations). Consequently, item N6 more strongly activates the high-dispersion category's cluster after learning. These effects are illustrated in the bottom panel of Figure 3. The ratio of cluster activations for stimulus N6 leads to a 70% probability of selecting the high-dispersion category, in close agreement with the empirical data.

Experiment 2

Both rational and mechanistic accounts captured Experiment 1's main finding—key border item N6 was assigned to the high-variance category during transfer. Given the elegance, soundness, and non-arbitrary form of the rational accounts, one might question the value of a mechanistic account that requires consideration of unobservable learning processes and category representations. To the contrary, we argue that the worth of mechanistic accounts lies in these considerations and that consequent deviations from rationality (even ostensibly minor ones) can lead to important insights into human behavior.

In Experiment 2, we test one prediction of the mechanistic model. Unlike the rational models, the mechanistic model updates its memory representation of each category in a local trial-by-trial fashion. The mechanistic model predicts that perceptions of category variability are based on trial-by-trial discrepancies between the current stimulus and the memory representation of the category (i.e., the position of the respective cluster). Thus, the mechanistic model predicts that humans' perception of category variability can be manipulated by altering the sequence of training items, whereas the rational models do not predict such effects.

In Experiment 2, members of one category appeared in an ordered fashion such that successively presented members did not vary much from each other. In contrast, members of the other category were presented in a random fashion, as were members of both categories in Experiment 1. In Experiment 2, the mechanistic model predicts that the discrepancy between the position of a category's cluster and the current stimulus will be smaller on average for the ordered category, and therefore humans should treat the random category as more variable and assign item N6 to it. This prediction is based on how cluster positions are updated in a local, trial-by-trial fashion. For the random category, the cluster position will fluctuate tightly around the true category mean, whereas in the ordered category the cluster position will smoothly track the periodic oscillations created by the ordering manipulation (see Figure 5), leading to smaller average discrepancies and a lower estimate of category variability. To foreshadow Experiment 2's results, the predictions of the mechanistic model held.

Methods

Forty-eight University of Texas undergraduates were tested. The procedure was the same as in Experiment 1 except for the line lengths and order of stimulus presentation. Stimuli ranged from 60 to 180 pixels in one category and from 260 to 380 pixels in the other. Adjacent items differed by 5 pixels, resulting in 25 items per category. Whether the ordered category had longer or shorter lines than the random category was counterbalanced across participants.

Every member of each category appeared exactly twice during training. The presentation order for this phase was determined by first generating a sequence for each category and then randomly interleaving these sequences in blocks of ten (five from each category sequence). The sequence for the ordered category was designed to reduce local variability. This sequence proceeded from the middle of the category distribution to the extreme (i.e., moving away from the category boundary), from the extreme to the category boundary (passing through the middle), and then from the boundary back to the middle of the category. More precisely, the sequence was generated by starting with the sequence O12, …, O1, O1, …, O25, O25, …, O13 and swapping each adjacent pair (excluding the first and last) with probability .5. Under this scheme the items closest to border item N6 are presented after the items furthest away, so a simple explanation from recency effects works against our hypothesis. The presentation order for the random category was random except for the first and last items, which were constrained to be R14 and R13, respectively (mirroring the ordered category). Figure 4 shows an example stimulus sequence.

The transfer stimuli consisted of lines of lengths 40 and 50 (novel items); 60, 90, 120, 150, and 180 (training items); 190, 200, 210, 220, 230, 240, and 250 (novel items); 260, 290, 320, 350 and 380 (training items); and 390 and 400 (novel items), with 220 as the critical border item N6. As in Experiment 1, subjects completed two blocks of transfer, each with a random presentation order.

Results and Model Fits

Participants were more likely to classify border item N6 into the random than the ordered category. As shown in Figure 5, averaged across the two transfer blocks, subjects assigned the border item to the random category with probability .80, which is significantly greater than chance, $t(47) = 6.19$, $p < .001$. In the first transfer block, more subjects (38 of 48) classified the border item in the random category than was expected by chance, exact binomial $p = .000062$ (two-tailed). The same pattern (39 of 48) was found in the second transfer block, exact binomial $p = 0.000015$ (two-tailed).

The mechanistic model was fit to Experiment 2's data using the same parameter values as those used in Experiment 1's simulation. As expected, the cluster mean for the ordered category tracked the stimuli, leading to lower average discrepancy between the cluster mean and each current stimulus, which in turn resulted in a lower variability for that cluster than the cluster for the random category. This local effect, shown in Figure 4, resulted in average standard deviations of 17.9 for the ordered category and 25.6 for the random category after learning. Consequently, item N6 more strongly activated the random category's cluster, leading to a 79% probability of selecting the random category, in close agreement with the human result.

General Discussion

In Experiment 1, people appeared sensitive to category variability and assigned a transfer item lying between two categories into the higher-variability category. This finding suggests natural accounts from both mechanistic and rational perspectives. These accounts largely converge, in that both assume people learn the mean and variability of each category and use that information to classify new items. One distinguishing and non-rational aspect of the mechanistic account is that estimates of category mean and variance are made in a trial-by-trial fashion. Instead of calculating an unbiased estimate of these quantities, the mechanistic model employs local learning rules that are driven by discrepancies between the memory representation of the category (i.e., the cluster) and the current stimulus.

This departure from rationality might seem modest, but it was the basis for a surprising prediction that was confirmed in Experiment 2. In Experiment 2, both categories had equal variance, but one category was ordered semi-regularly such that differences between stimuli on successive trials were small. The mechanistic model predicted that this ordering would create an illusion of low variability for the ordered category, as the discrepancy between each presented stimulus and the current category representation was relatively small. Accordingly, human subjects assigned the border item at transfer into the randomly ordered category. Overall, these empirical and modeling results suggest that people estimate variability by making incremental adjustments to memory representations based on local comparisons. These results also suggest that consideration of mechanistic models, with their accompanying processes and representations, is a fruitful research strategy, particularly when departures from rationality are considered.

One persistent criticism of the mechanistic approach is that multiple mechanisms can give rise to the same behavior (Townsend, 1974). Proponents of the rational approach argue that grounding models in the structure of the environment provides additional constraints. Although incorporating additional constraints is desirable, we find the claims that there are privileged and unambiguous facts about the environment to be dubious. Any environment of sufficient complexity can be characterized in a number of different ways. Assumptions about what information people monitor, the dynamics of the environment, and associated rewards can vary. When a rational analysis fails, these assumptions are altered until the desired result is achieved (as in Step 6 of Anderson's, 1990, rationality framework).

Importantly, these attacks on mechanistic accounts ignore the substantial constraints that such a perspective provides. Mechanistic accounts are not made in a theoretical vacuum but are informed by existing models and behavioral findings. Current thinking on the nature of our cognitive architecture (e.g., capacity-limited working memory) provides grounding, and the successes and failures of related models provide lessons for future models. This was the case in formalizing the mechanistic model presented here, in light of the substantial evidence for similarity-based representations and error-driven learning, coupled with the specific failures of prototype and exemplar models in explaining Experiment 1. In practice, the mechanistic approach may offer more constraints than the rational approach's “first principles” orientation, which emphasizes de novo analysis of the current task and environment for each application.

We are not suggesting that there is not a rational account of Experiment 2's results. There are likely an infinite number of possible rational explanations. For example, a rational account that assumes categories in the environment steadily drift could be made consistent with Experiment 2's results. Along these lines, Elliott and Anderson (1995) present a rational model that makes these assumptions in regards to estimating a category’s mean. Interestingly, their model’s disproportionate weighting of recent items leads it to predict the opposite patterns of findings from what we observe in Experiment 2.

The key meta-scientific question is whether successful rational explanations would come to the forefront prior to specifying Experiment 2's design. Following Experiment 1, we specified the most straightforward and readily suggested rational and mechanistic accounts. Focusing on the rational explanation of Experiment 1 would not have led to Experiment 2, whereas considering *how* the mechanistic model accounted for Experiment 1's results did motivate Experiment 2.

One possibility is that rational explanations, while illuminating and satisfying, largely serve as just-so stories that are constructed after interesting behavioral findings present themselves. According to this view, rational analyses are more likely to follow from mechanistic explanations than vice versa. Perhaps one example of this progression is from the RULEX (Nosofsky, Palmeri, & McKinley, 1994) model of hypothesis generation and testing to Boolean complexity (Feldman, 2000). RULEX specifies how people search for Boolean rules by beginning with simple rules and progressing toward more complex rules when simple rules fail. Boolean complexity preserves many of these insights, albeit in a more abstract form that does away with RULEX's proposed search and memory processes. Instead, Boolean complexity offers a well-formulated metric that is derived through a rational analysis.

Although our discussion has been provocative and heavily tilted in favor of mechanistic approaches, we do not wish to suggest that rational analysis does not have its place. Here, we suggest that mechanistic models can guide rational analyses. Likewise, rational analyses can guide the development of mechanistic models. A rational analysis can uncover the principles that mechanistic models approximate and bring into focus how the mechanistic model deviates from rationality. Experiment 2's design was motivated by such considerations. In addition, consideration of what environmental assumptions would rationally justify behaviors exhibited by mechanistic models can provide insight into our cognitive environment, such as the idea that real categories drift over time as suggested by a post hoc rational analysis of Experiment 2. As a field, we are likely to make progress when intellectual effort is devoted to both approaches. Given the recent tilt toward rational approaches, we would like to end by encouraging the field not to shy away from mechanistic explanations. If the main question we are trying to answer is how the mind works, we should not fear directly addressing this question by developing and evaluating mechanistic models.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*, 1165–1175.

Elliott, S. W., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 4, 815-836.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 234–257.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767-773.

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & Cognition*, *33*, 289–302.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, *111*, 309–332.

Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 301–321.

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

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Ono, H. (1967). Difference threshold for stimulus length under simultaneous and nonsimultaneous viewing conditions. *Perception & Psychophysics*, *2*, 201–207.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York: Cambridge University Press.

Sakamoto, Y., Matsuka, T., & Love, B. C. (2004). Dimension-wide vs. exemplar-specific attention in category learning and recognition. In M. Lovett, C. Schunn, C. Lebiere, &

P. Munro (Eds.), *Proceedings of the 6th International Conference of Cognitive Modeling* (pp. 261–266). Mahwah, NJ: Lawrence Erlbaum Associates.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (p. 133-186). Hillsdale, NJ: Lawrence Erlbaum Associates.

Author Note

This work was supported by AFOSR FA9550-04-1-0226 and NSF CAREER #0349101 to B.C.L., NIMH NRSA F32-MH068965 to M.J., and ONR N00014-05-1-00632 to Y.S. Correspondence concerning this research should be addressed to brad_love@mail.utexas.edu.

Footnote

[1] All of these models assume that category members are distributed according to the Gaussian (i.e., normal) distribution. This common choice is motivated by a variety of considerations ranging from the nature of noise in the nervous system to the general structure of categories in our environment. Rational models can be formulated using other distributions when such distributions better conform to the structure of a particular domain (cf. Griffiths & Tenenbaum, 2006).

Figure Captions

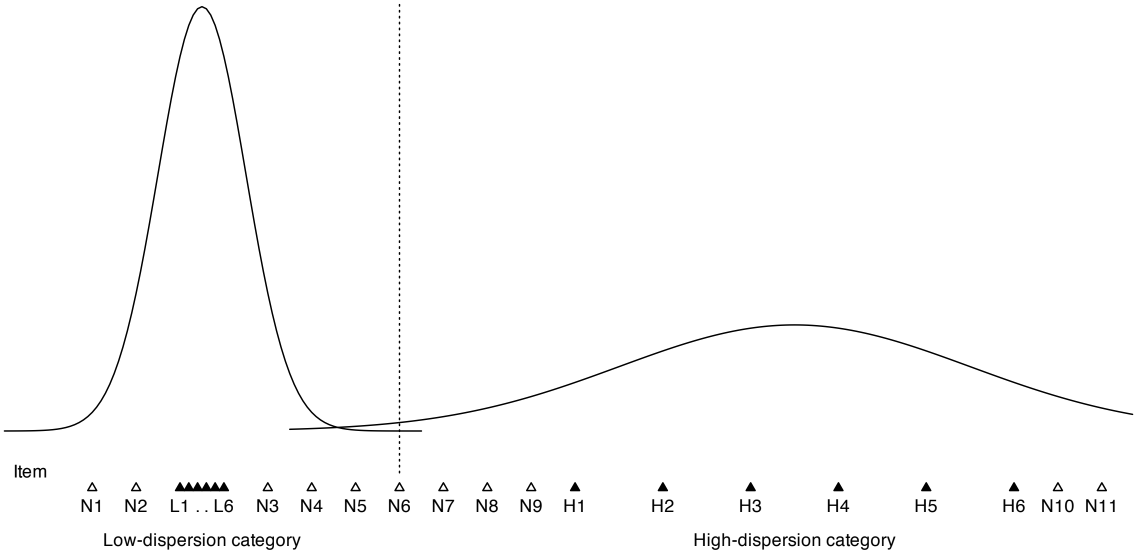
Figure 1. The design of Experiment 1. Dark triangles (L1–L6 and H1–H6) represent studied items and light triangles (N1–N11) represent novel items that did not appear during learning. The item lengths are spaced to scale. Item N6 is exactly midway between the nearest studied members (L6 and H1) of the low- and high-dispersion categories. Each studied item in the low-dispersion category differs from its nearest neighbor by 2 pixels, whereas each studied item in the high-dispersion category differs from its nearest neighbor by 20 pixels. To eliminate possible influences of absolute line length on performance, whether the high-dispersion category had longer (condition C1) or shorter lines (condition C2) than the low-dispersion category was counterbalanced between subjects. The two density functions are illustrative of the category representations developed by both rational and mechanistic models when applied to this task.

Figure 2. The probability of subjects' classifying each stimulus item as a member of the high-dispersion category during the transfer phase of Experiment 1. Training items are shown as dark triangles; novel items are shown as light triangles. Item N6 is midway between the nearest studied members (L6 and H1) of the low- and high-dispersion categories. Items are not spaced to scale (see Figure 1 for the physical scale).

Figure 3. The activations of the clusters encoding the low- (cluster L) and high-dispersion (cluster H) categories in the mechanistic model are shown for each stimulus item in Experiment 1. The top panel shows that the variability for each cluster is equal before learning. An arrow indicates the number of standard deviations from the mean of each cluster to border item N6. Before learning, the border item is fewer standard deviations from cluster L's than cluster H's center, leading to greater activation and higher response probability for the low-dispersion category. The bottom panel shows that the opposite pattern arises after learning, due to the tightening of cluster L and the widening of cluster H (which make each cluster relatively more responsive to its category's members). After learning, the border item is more likely to be assigned to the high-dispersion category.

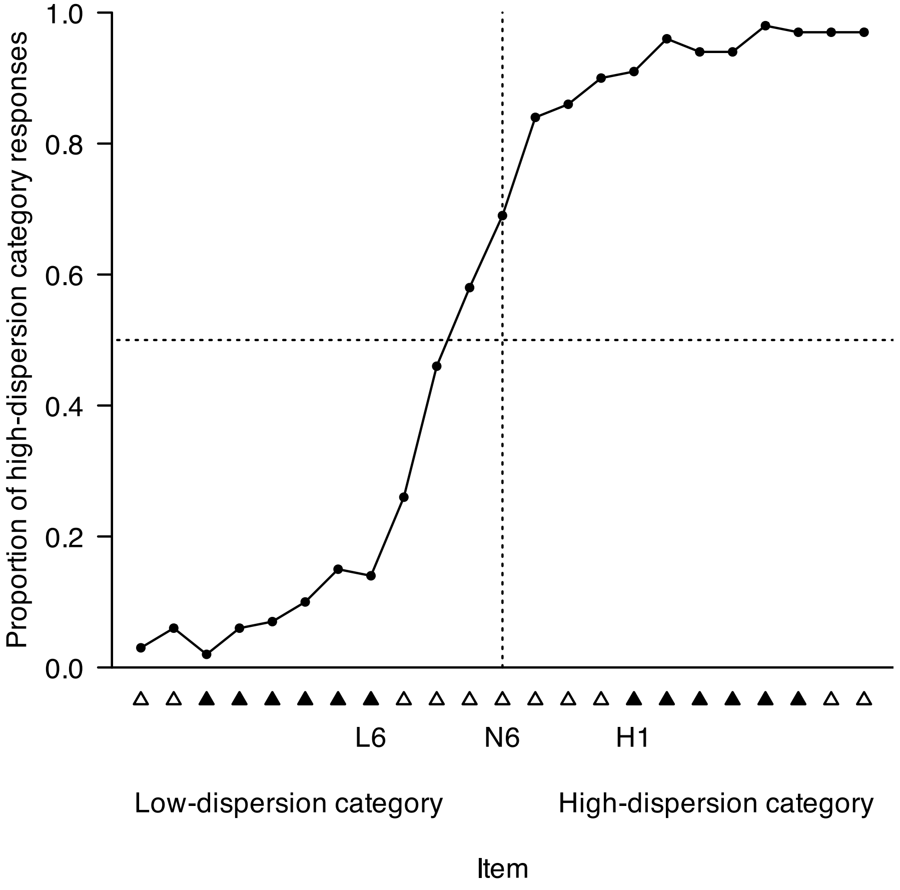
Figure 4. Cluster position (i.e., estimated category mean) learning over a typical simulation of the mechanistic model on Experiment 2. The horizontal axis denotes stimulus length and the vertical axis captures the learning trial sequence. Each training stimulus is depicted by a solid triangle. Evolving cluster positions are shown by the solid lines. The cluster position for the ordered category follows the trajectory of the learning items. In comparison to the random category, this tracking leads to smaller differences between each stimulus and the current cluster position. Because of these smaller discrepancies, the model learns a lower variability for the ordered category and assigns border item N6 to the random category, in agreement with human subjects.

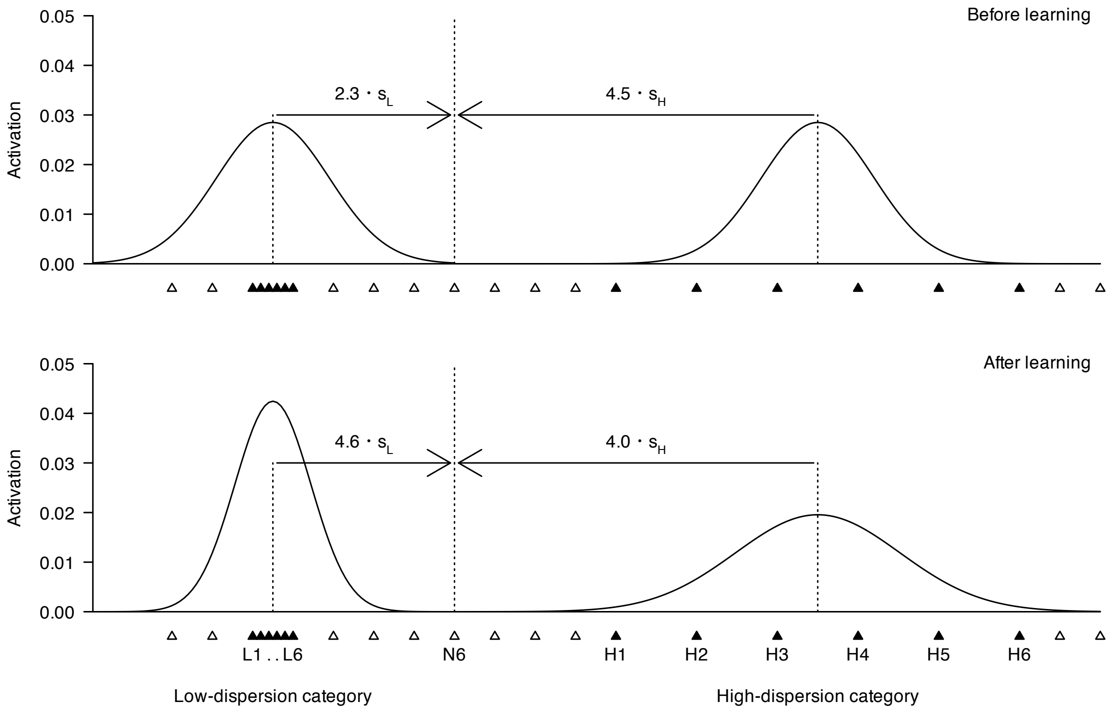
Figure 5. The probability of subjects' classifying each stimulus item as a member of the random category during the transfer phase of Experiment 2. Learning items are shown as dark triangles; novel items are shown as light triangles. Item N6 is midway between the nearest studied members (O25 and R1) of the ordered and random categories (and also midway between the prototypes of the two categories).

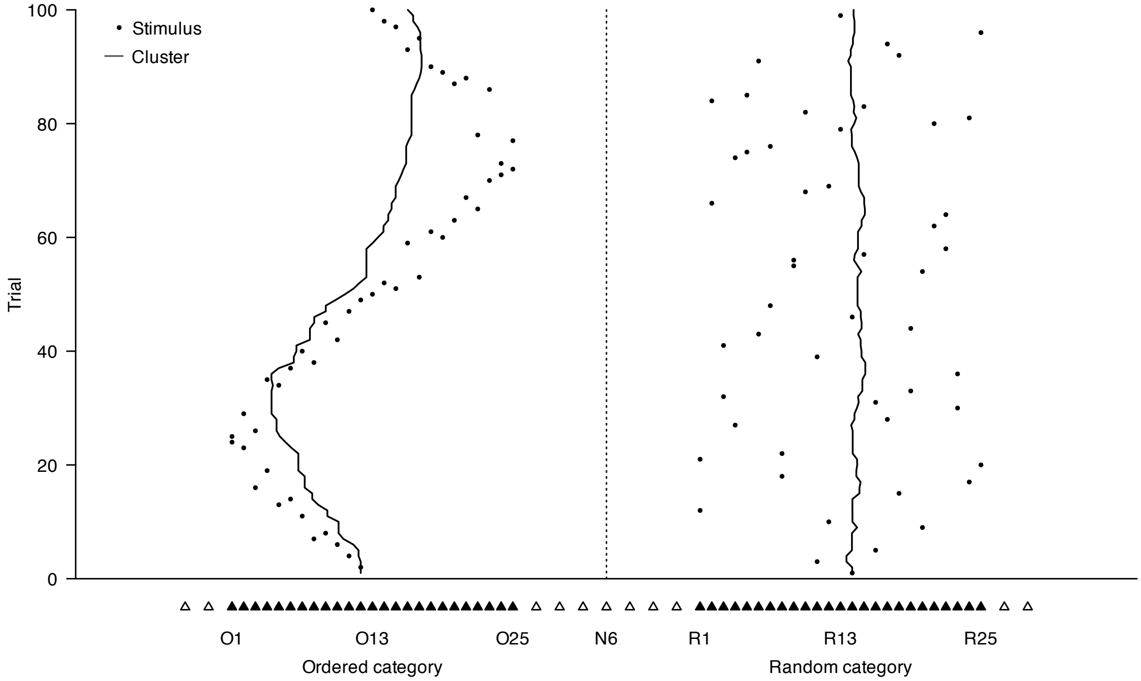


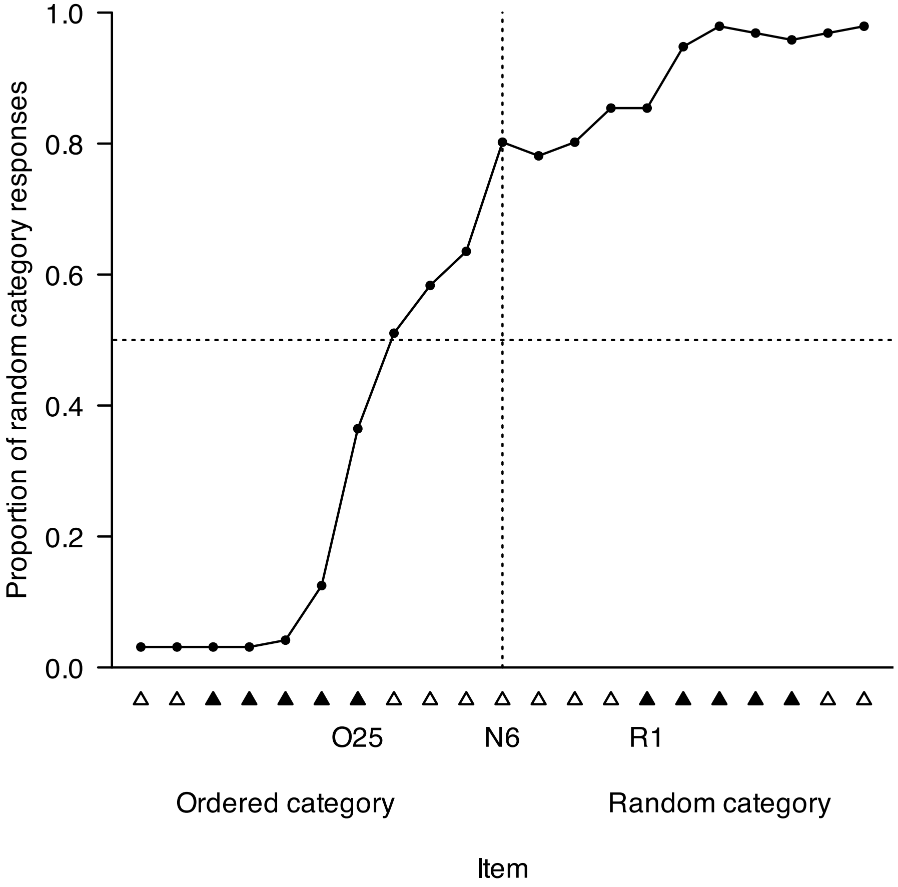
Length in pixels (100 pixels = 33.25 mm)

C1:	120	130	140. .150	160	170	180	190	200	210	220	230	250	270	290	310	330	340	350
C2:	260	250	240. .230	220	210	200	190	180	170	160	150	130	110	90	70	50	40	30


Proportion of high-dispersion category responses
1.0
0.8
0.6
0.4
0.2
0.0
L6
N6
H1
Low-dispersion category
High-dispersion category
Item


Before learning
0.05
0.04
0.03
0.02
0.01
0.00
Activation
2.3 · s_L
4.5 · s_H
After learning
0.05
0.04
0.03
0.02
0.01
0.00
Activation
4.6 · s_L
4.0 · s_H
L1 . . L6
N6
H1
H2
H3
H4
H5
H6
Low-dispersion category
High-dispersion category


100
80
60
40
20
0
Trial
• Stimulus
— Cluster
O1
O13
O25
N6
R1
R13
R25
Ordered category
Random category


Proportion of random category responses
1.0
0.8
0.6
0.4
0.2
0.0
△ △ ▲ ▲ ▲ ▲ ▲ △ △ △ △ △ △ △ ▲ ▲ ▲ ▲ ▲ △ △
O25
N6
R1
Ordered category
Random category
Item