Belief Revision in Models of Category Learning

Evan Heit
Department of Psychology
Northwestern University
heit@nwu.edu

Abstract

In an experiment, subjects learned about new categories for which they had prior beliefs, and made probability judgments at various points during the course of learning. The responses were analyzed in terms of bias due to prior beliefs and in terms of sensitivity to the content of the new categories. These results were compared to the predictions of four models of belief revision or categorization: (1) a Bayesian estimation procedure (Raiffa & Schlaifer, 1961); (2) the integration model (Heit, 1993, 1994), a categorization model that is a generalization of the Bayesian model; (3) a linear operator model that performs serial averaging (Bush & Mosteller, 1955); and (4) a simple adaptive network model of categorization (Gluck & Bower, 1988) that is a generalization of the linear operator model. Subjects were conservative in terms of sensitivity to new information, compared to the predictions of the Bayesian model and the linear operator model. The network model was able to account for this conservatism, however this model predicted an extreme degree of forgetting of prior beliefs compared to that shown by human subjects. Of the four models, the integration model provided the closest account of bias due to prior beliefs and sensitivity to new information over the course of category learning.

Imagine that you are an American traveling in Europe for the first time. Until now, your concepts of people, locations, and things in European cities have been largely shaped through media images rather than direct experience. For example, suppose that your concept of people in City P includes the strong prior belief that these people tend to be unfriendly to Americans. To be specific, you might expect that on 90% of your encounters with people in City P, the person you meet will be rude or unfriendly. Now say that during the first day of your visit, you meet ten citizens of P. To your delight and surprise, only three of them act unfriendly to you. Your expectations about people in City P may have been derived from an inaccurate stereotype. Clearly, your concept of these people must be revised in light of these new observations. But how much revision should take place? When you travel the next day, will you expect the majority of people to be unfriendly or friendly? Say that on the next day, you meet ten more people, and again, three people are unfriendly. How do you put together your prior knowledge, the first day's observations, and the second day's observations?

Belief revision is an important task that people face often, even when they are not traveling the world. Whenever people learn new concepts, they may bring to bear their previous expectations and theories (Murphy & Medin, 1985; Murphy, 1993). Typically there will be some discrepancy between prior knowledge and what is observed about a new category, otherwise there would be nothing to learn. Therefore, category learning may be considered as a kind of belief revision. Furthermore, categories in the world can change over time, so beliefs about these categories must be updated periodically. For example, improvements in technology have led to changes in people's concepts of telephones (Elliott & Anderson, in press).

The experiment in this paper addresses the dynamics of category learning, in which performance depends on prior knowledge and new observations. Subjects had initial beliefs about persons in a fictional place referred to as *City W*, then gradually revised their concepts as they observed descriptions of persons in City W. The results are considered in terms of four computational models of belief revision.

Method

Overview. First, the subjects' initial beliefs about various categories of people in a new city, e.g., *shy people*, *joggers*, were assessed. Then the subjects observed descriptions of people in City W. Some descriptions were congruent with prior knowledge, such as a shy person who avoids parties, and some descriptions were incongruent, such as a jogger without expensive running shoes. This observation phase was interrupted periodically as subjects were asked to make probability judgments about transfer stimuli. Overall, the procedure was similar to that of Heit (1994), except that the subjects' changing beliefs were assessed a total of five times over the course of learning.

Stimuli. Each subject saw training examples derived from five couplets of descriptive terms (see Table 1; the complete pool is in Heit, 1994). Each couplet of four features was comprised of two pairs of opposites or complements. For example, not shy is the complement of shy, and does not attend parties often is the complement of attends parties often. The first and third item in each couplet were congruent with each other (e.g., shy and does not attend parties often), likewise the second and fourth item were congruent. The first and fourth items, as well as the second and third items, were incongruent (e.g., shy and attends parties often). The stimuli were pre-tested on other subjects to validate this manipulation of prior knowledge (see Heit, 1994).

Table 1: Feature Couplets (Examples)

Shy / not shy
Does not attend parties often / attends parties often
Jogs regularly / does not jog regularly
Owns expensive running shoes / does not own expensive running shoes
Travels two or more times per year / travels less than two times per year
Has frequent flyer number / does not have frequent flyer number
Watches more TV than average / watches less TV than average
Reads books less than average / Reads books more than average
Generous / not generous
Donates to charity / does not donate to charity

Each training example was a description of a person, in terms of two features from a couplet. A pairing of two features was either congruent or incongruent. The five couplets were assigned randomly for each subject to the following structure: one couplet had 0% congruent pairings, one couplet had 25% congruent pairs, one couplet was 50% congruent, one was 75% congruent, and one was 100% congruent. For example, when the shyness-parties couplet was assigned to the 100% congruent condition, all the shy people did not attend parties often and all the non-shy people did attend parties often.

Procedure. At the beginning of the experiment, subjects were told that they would see descriptions of persons living in City W, a city located in Illinois. The procedure followed a test-study-test-study-test-study-test-study-test sequence. In the first test block, subjects' prior beliefs about people in City W were assessed, presumably reflecting their general knowledge about people in Illinois. Within a study block, the training examples were presented individually, in a random order, about every 3.5 seconds. In each study block, subjects were presented with forty training examples, eight per couplet. In effect, subjects were given four members of each category during a study block. For example, subjects would see four descriptions of shy persons in each study block. Each of the four study blocks was followed by a test block. Thus, subjects were tested after they had observed 0, 4, 8, 12, and 16 members per category.

In each test phase, subjects made 20 conditional probability estimates. These questions were worded as follows:

Consider a person from City W with the following characteristic: x
How likely is it that this person would also have this characteristic? A,

where x and A were two features. Subjects responded on a scale from 0% to 100%. The test stimuli for each block had a two factor design: (1) whether the two features were congruent or incongruent with each other; and (2) the conditional probability of presentation during the study phase, which was 0%, 25%, 50%, 75%, or 100%. Eight test questions were derived from each couplet, thus there were 40 possible test questions. In each test phase, 20 of these questions were chosen randomly and asked of the subject.

Subjects. Forty-two Northwestern University undergraduates participated.

Results

The average responses at different points during the experiment are shown in the first column of Figure 1. The top panel shows the initial responses, before any training had begun. Here, subjects clearly were influenced by their prior knowledge, as indicated by the higher judgments for congruent test questions (e.g., how likely a jogger is to own expensive running shoes) compared to incongruent test questions (e.g., how likely a shy person is to attend parties often). There was no influence of observed proportion of stimuli co-occurring, because at this point the subjects had not observed any training stimuli. (For the top panel, the observed proportions were defined for the experiment but not known to the subjects.) The lower panels in this column, corresponding to category sizes 4, 8, 12, and 16, show that subjects did revise their beliefs as they observed people in City W. Two concepts are critical for understanding these trends.

First, it is useful to consider subjects' *bias*, that is, the direct influence of prior knowledge about these categories. Bias may be measured in terms of the difference between congruent and incongruent judgments, at a given level of observed proportion. For example, a subject with no bias would show zero difference between congruent and incongruent lines. The second consideration is the subjects' *sensitivity* to what they observed. That is, how well do the estimated proportions reflect the observed proportions of category membership in City W? Sensitivity can be measured in terms of the slope of the lines in Figure 1. A zero slope indicates no sensitivity to observed proportion, and higher slopes approaching one indicate greater sensitivity.

Now, looking at the panels in the first column of Figure 1 from top to bottom, two trends are apparent. First, with more experience, subjects became less biased by prior knowledge; the lines tend to converge. Second, subjects became more sensitive to the observed data with more experience; the slopes increase. (Also see Table 2.)

Statistical analyses supported these observations. There was a main effect of congruent versus incongruent test question, indicating that subjects gave higher judgments for congruent questions, $F(1,41)=83.0$, $p<.001$. The congruent versus incongruent factor exhibited an interaction with test block, such that the difference between congruent and

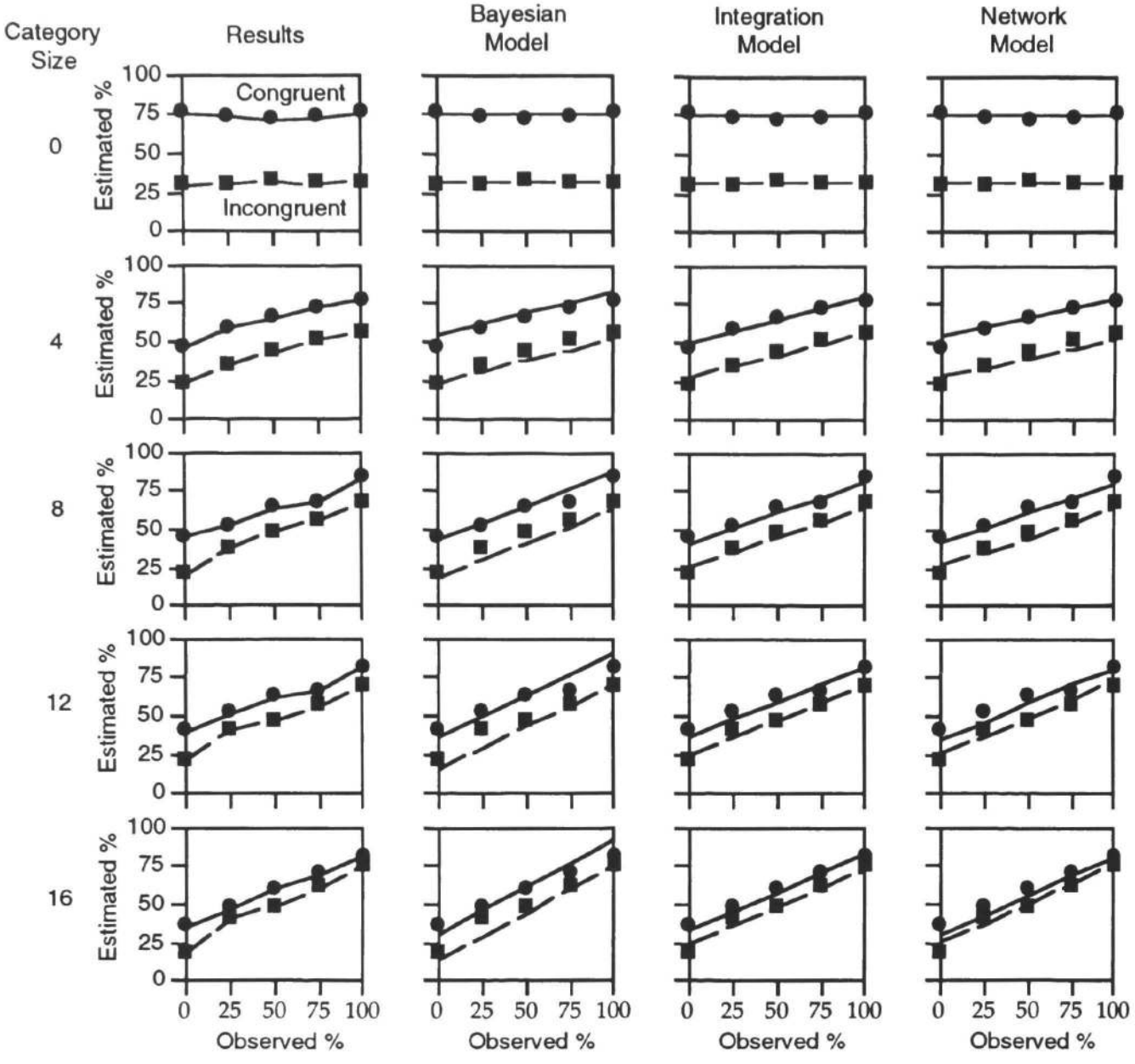

Figure 1: Results and model predictions.

incongruent questions diminished with more items studied, $F(4, 164)=37.0$, $p<.001$. There was also a main effect of objective probability, $F(4, 164)=70.2$, $p<.001$, indicating that subjects' judgments were sensitive overall to what they observed. This effect of objective probability interacted with testing block, $F(16, 656)=23.0$, $p<.001$, such that subjects were more sensitive to objective probability as more items were studied.

Model Evaluations

Four models were considered as accounts for these results. The first two, the Bayesian revision model and the linear operator model, were included because they are classic models of learning and judgment. The integration model is an exemplar model of categorization that generalizes the Bayesian model. Also, a simple connectionist network model, a generalization of the linear operator model, was considered.

Bayesian and Linear Operator Models

The Bayesian revision model provides a means for combining a prior belief about a statistical parameter with new observations (Raiffa & Schlaifer, 1961). This model is especially important because within the framework of Bayesian statistical theory it is taken to be a normative procedure. Therefore, comparing the results of this experiment to the predictions of the Bayesian model gives some perspective on whether subjects were behaving optimally. The Bayesian formula for revising an estimate of a proportion is shown in Equation 1.

$$P_N = \frac{Np + Gq}{N + G} \quad (1)$$

In this equation, P_N is the estimated proportion of items with description x that belong to category A, after N observations have been made. The variable q represents the prior estimate of this proportion (before any observation), and G indicates the strength of this prior belief. The variable p is the proportion of the new observations that belong to category A. As N increases, the estimate depends more on the observed proportion, p, and less on the prior belief, q.

The linear operator model is derived from classic mathematical learning theory (Bush & Mosteller, 1955), and this model has been applied to numerous results in learning and probability judgment (see Bower & Heit, 1992). Furthermore, this formula is useful for calculating a running estimate of a proportion using an anchor-and-adjust procedure (Busemeyer, 1991). This model is shown in Equation 2.

$$P_N = P_{N-1} + \beta(d_N - P_{N-1}) \quad (2)$$

Note that P_0 is set to q, the prior belief about this proportion. The indicator variable d_N refers to what is observed on trial N; it is assigned a value of 1 when the observed item is in category A and a value of 0 when the observed item is not in category A. Note that the expected value of d_N in Equation 2 is equivalent to the proportion p in Equation 1. Finally, β refers to the learning rate, between 0 and 1. The estimated proportion after observation N is the previous estimate, from trial $N - 1$, plus a correction, determined by difference between what is observed, d_N, and the previous estimate, P_{N-1}. In the asymptote, P_N will approach p, the proportion of observed items in category A.

Note that each of these models has two free parameters.[1] The value of q, the prior estimate of the proportion, was estimated for each model from the responses on the first block of test trials, before any observations had taken place. The best-fitting value of, q, derived algebraically, was .72.

To estimate the other parameters, the two models were fitted to the average responses on the 50 test questions. The values of G and β in the two models were estimated by simulating each model with various parameter values, and searching through the parameter space with the criterion of minimizing the root mean square error of prediction over the 50 judgments. For the Bayesian model, when G had the value of 10.08, the root mean square error (RMSe) of the model was .0504. For the linear operator model, when β had the value of .064, the RMSe of the model was .0563. (Because trial order has a slight effect on the predictions of the linear operator model, this model was simulated with 100 different random orders of trials, and the predictions were averaged.) In terms of the RMSe performance measure, the two models are close, but the Bayesian model is slightly better.

The best-fitting predictions of the Bayesian models are shown as the lines in the second column of Figure 1, overlaid on the data points. (The predictions of the linear operator model are not shown, but they are similar to the Bayesian model.) The models capture the qualitative pattern of belief revision: With more observations of category members, the models predict that bias decreases and sensitivity increases. However, the subjects' responses were much less sensitive than what is predicted by the models. This finding is evident in the second column from comparing the slopes of the lines to the slopes of the data points--the lines have steeper slopes, indicating greater sensitivity for the models. Another way of stating this result is that, compared to the normative Bayesian model as well as the linear operator model, subjects were *conservative* in terms of sensitivity.

The subjects' conservatism might derive from additional details of processing, such as memory confusions or a lack of sensitivity in the response scale, not captured by these models. The next two models to be described are similar to the Bayesian and linear operator models, except for additional processing assumptions.

Integration and Network Models

The integration model (Heit, 1993, 1994) is an exemplar model of categorization that has already been applied to several categorization experiments where subjects were influenced by pre-experimental knowledge. The critical claim of the integration model is that when people learn about a new category, they are influenced by prior examples from other, related categories. Information from prior examples and from new observations is simply summed together. For example, in learning about shy people in City W, subjects would be influenced by memories of shy people from other places as well by actual observations of people in City W. Such transfer of memories from one source or context to another has been described and documented by Johnson, Hashtroudi, and Lindsay (1993).

For the present experiment, in which subjects predicted a category given a single feature, the integration model is a generalization of the Bayesian revision model. The integration model is described by Equation 3 (see Heit, 1994).

$$P_N = \frac{Np + Gq + sN(1-p) + sG(1-q)}{N + G + sN + sG} \quad (3)$$

The new variable in this equation is s, which measures the degree of confusions in memory (see Medin & Schaffer, 1978). The value of s may range from 0 to 1, with greater values indicating poorer feature memory. Note that when $s = 0$, Equation 3 is equivalent to Equation 1 for the Bayesian model. In the integration model, G is interpreted as a number of prior examples retrieved for a given category A, and q is the proportion of prior examples with description x.

[1] The average of all the responses in this experiment was 53%; however, each of the four models to be considered predicts an average response of 50%. The discrepancy seems to be due to a slight lack of calibration by the subjects (see Heit, 1994; Wallsten & Gonzalez-Vallejo, 1994). To compensate for this difference, a correction of .03 was added to every prediction of each model.

The final model to be considered is a simple connectionist network model proposed by Gluck and Bower (1988). This model learns direct associations between input units and output units using the least-mean-square (LMS) learning rule. This model is particularly appropriate because subjects predicted single variables from a single cue. The network model is quite similar to the linear operator model in Equation 2, with a few exceptions. First, the training signals (d_n) as well as the prior belief (q) range from -1 to 1 instead of 0 to 1. Second, the output of the linear operator model is passed through a logistic activation function, $P_N = 1 / (1 + \exp(-\theta O_N))$. Here, O_N refers to the value obtained from Equation 2, and θ is a scaling parameter. Increasing values of θ indicate greater overall sensitivity in judgments.

Each model has three free parameters. To make the fits of the integration and network models comparable to other models, these models were constrained to fit the initial set of judgments as closely as possible. After some algebraic manipulation, the models were constrained as follows. For the integration model, q must be equal to $.5 + ((.22)(1 + s)) / (1 - s)$ to fit the initial judgments. For the network model, $q = .94 / \theta$.

With those constraints set, the free parameters were estimated. For the integration model, when the q parameter was .85, G was 4.73, and s was .23, the RMSe of the model was .0261, about half the error of the Bayesian model. For the network model, when the value of q was .73, β was .12, and θ was 1.29, the RMSe of the model was .0364, intermediate between the integration model and the other two models. The predictions of these two models are shown as the lines in third and fourth columns of Figure 1.

With the additional free parameters allowing some degree of memory confusions or lack of responsiveness, the models now give good accounts of subjects' sensitivity over the course of learning. Note that in the third and fourth columns, unlike the second column, the slopes of the lines (the model predictions) are quite close to the slopes of the data points. The integration model also gives an excellent account of how subjects' biases due to prior knowledge, in terms of the difference between congruent and incongruent lines, change over the course of learning. In contrast, the network gives a poor account of bias due to prior knowledge. The network model predicts that the initial biases will be nearly forgotten near the end of learning, i.e., the lines nearly converge in the bottom two panels of the fourth column. In contrast to this prediction, subjects still showed substantial bias towards the end of learning. This model's rapid forgetting of earlier beliefs is similar to the phenomenon of catastrophic interference in more complex networks (Ratcliff, 1990).

Why does the integration model gives a better account of the prior knowledge bias than the network model? The integration model assumes that prior examples and previous observations have a persistent influence, even as new examples are observed. As a learner accumulates more information, the marginal influence of each additional observation decreases. This can be seen by rewriting Equation 1 as a difference equation analogous to Equation 2. (Equation 3 can also be rewritten as a difference equation to make a similar point, but the resulting equation is more unwieldy.)

$$P_N = P_{N-1} + \frac{1}{G+N}(d_N - P_{N-1}) \qquad (4)$$

What is critical in Equation 4 is that the learning rate, $1 / (G + N)$, decreases as more items are observed, i.e., as N increases. In contrast, in the learning rule for the network model in Equation 2, the estimate is revised at a fixed rate, β, regardless of the number of previous observations. Towards the end of the experiment, the network model was revising too quickly compared to the subjects. It should be possible to improve the performance of the network model by additionally assuming that the learning rate decreases over the course of the experiment.

Conclusion

These results are consistent with previous results on probability revision, such as the classic urns-and-balls problems reviewed by Edwards (1968). In those studies, subjects were also conservative compared to a Bayesian model, in terms of sensitivity to observed proportions. The present experiment differs from those older studies in that subjects' prior beliefs were derived from real-world social knowledge (e.g., about shy people). In related research, Elliott and Anderson (in press) examined the learning of categories that change over the course of an experiment. Elliott and Anderson also found that an exemplar model gives a better account of belief revision than a network model. In addition, they found evidence for forgetting of earlier observations, so that an exemplar model with assumptions of memory decay performed even better. (In

Table 2: Summary of sensitivity and bias for experimental results and model predictions.

	Sensitivity (Slope)					Bias				
Category Size	Results	Bayesian Model	Lin. Op. Model	Integ. Model	Network Model	Results	Bayesian Model	Lin. Op. Model	Integ. Model	Network Model
0	.01	.00	.00	.00	.00	.43	.43	.43	.43	.43
4	.32	.29	.23	.29	.23	.22	.32	.34	.23	.27
8	.43	.44	.41	.40	.38	.17	.25	.26	.16	.16
12	.45	.54	.55	.45	.46	.14	.20	.20	.12	.10
16	.51	.61	.65	.49	.50	.10	.17	.15	.09	.06

contrast, the integration model does not implement memory decay.) Elliott and Anderson's work is well suited to investigate forgetting because their categories changed over the course of learning, unlike the present experiment in which the categories did not change. Yet their procedure may have encouraged subjects to strategically ignore early observations, because using the older observations would lead to incorrect predictions. So what appears to be forgetting may also reflect some discounting of old information.

In summary, the present results show how people's concepts initially are influenced by prior beliefs and are revised gradually as new category members are observed. This process of belief revision can be described in terms of the integration model (Heit, 1993, 1994). According to this model, when people learn about a new category, they retrieve prior examples from related categories as well as accumulate examples that they actually observe for the category. At a general level, the predictions of the integration model are similar to those of other models, but at a more detailed level the integration model gives a more successful account of the course of learning and the relation between sensitivity and bias due to prior knowledge. (See Table 2 for a summary of the results and the models' predictions in terms of sensitivity and bias.) The detailed model comparisons suggest two additional principles that are central to the integration model's ability to fit these results: (1) allowing some degree of memory confusions and (2) persistent influence of previous beliefs such that the learning rate decreases as more knowledge accrues.

The simple nature of this experiment, in which subjects predicted category labels from information about single features, was useful in distinguishing among these models. In future research, it would be interesting to compare the integration model to more complex connectionist networks (e.g., Choi, McDaniel, & Busemeyer, 1993) for categorization experiments in which subjects are influenced by prior knowledge but learn about more complex multidimensional stimuli.

Acknowledgments

This research was supported by NIMH Grant 1 F32 MH10069 and NSF Grant 91-10245. I am grateful to Douglas Medin and Gordon Bower for discussions of this research. After August 1, 1995, address correspondence to Evan Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom.

References

Bower, G., & Heit, E. (1992). Choosing between uncertain options: A reprise to the Estes scanning model. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 21-43). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory, Volume I: Cognition.* (pp. 187-215). Hillsdale, NJ: Erlbaum.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.* New York: Wiley.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition, 21,* 413-423.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Elliott, S. W., & Anderson, J. R. (in press). The effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Heit, E. (1993). Modeling the effects of expectations on recognition memory. *Psychological Science, 4,* 244-252.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1264-1282.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

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

Murphy, G. L. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173-200). London: Academic Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Boston: Harvard University, Graduate School of Business Administration.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285-308.

Wallsten, T. S., & Gonzalez-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490-504.