JOURNAL OF MATHEMATICAL PSYCHOLOGY 39, 164–178 (1995)

A Comparison of Learning Models*

DANIEL FRIEDMAN, DOMINIC W. MASSARO, STEPHEN N. KITZIS, AND MICHAEL M. COHEN

University of California, Santa Cruz

We investigate learning in a probabilistic task, called "medical diagnosis." On each trial, a subject is presented with a stimulus configuration indicating the value of four medical symptoms. The subject responds by guessing which of two diseases is present and is then given feedback about which disease was actually present. The feedback is determined according to fixed conditional probabilities unknown to the subject. We test a normative Bayesian model as well as simple variants of well-known psychological models including the Fuzzy Logical Model of Perception, an Exemplar model, a two-layer Connectionist model and an ALCOVE model. Both the asymptotic predictions of these models (i.e., predictions regarding behavior after it has stabilized and learning is complete) and predictions of trial-by-trial *changes* in behavior are tested. The models are tested against existing data from Estes *et al.* (1989, *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 556–571) and new data from medical diagnosis tasks that include not only asymmetric but also symmetric base rates. Learning was observed in all cases in that subjects tended to match the objective probabilities of the symptom configurations more closely in later trials. All of the descriptive models give a more accurate account of performance than the normative Bayesian model. Relative to a benchmark measure, however, none of these models does an especially good job of characterizing asymptotic performance or the learning process. We suggest that future experiments should address individual performance, rather than group learning curves.

INTRODUCTION

Behavior changes with experience. An individual takes an action, observes some of the consequences, and perhaps later takes a different action in similar circumstances. The purpose of the present paper is to investigate such behavioral changes, which we regard as the manifestation of *learning*. We are interested in several questions, both normative and descriptive. Does behavior converge to an optimum? If so, is convergence as rapid as possible given available information? If not, can behavioral changes be described reasonably well by any simple model?

These and similar questions can be asked in quite general contexts. In this paper we confine our attention to a particular laboratory version of a "medical diagnosis" task. This task has been addressed in previous work on learning (Gluck & Bower, 1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1991) and is simple but nontrivial. After a short review of research in probabilistic learning situations, we describe the medical diagnosis task, the analytic framework we use, the question of optimality, and the formalization of the models for this task.

We base our analysis on five models, the normative (optimal) Bayesian model, the Fuzzy Logical Model of Perception (FLMP), the two-layer Connectionist model (CMP), an Exemplar model, and a recent exemplar-based network model (ALCOVE). The asymptotic predictions and learning predictions of these models are all presented in the context of the medical diagnosis task in order to keep the exposition self-contained and to fix ideas and establish notation.

Next we turn to the data. In addition to existing data from Estes *et al.* (1989), we analyze new data from simplified medical diagnosis tasks that include not only asymmetric but also symmetric base rates. We find evidence that subjects do learn in the sense that response frequencies to the symptom configurations approach objective probabilities more closely on later trials than on earlier trials. All of the descriptive models give a more accurate account of performance than the normative Bayesian model. Relative to a benchmark measure, however, none of these models does an especially good job of characterizing asymptotic performance or the learning process. In the last section we summarize our findings. Our results raise more questions than they answer, so we conclude by suggesting directions for more definitive work.

Correspondence and reprint requests should be sent to Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064.

* The research reported in this paper and the writing of the paper were supported in part by grants from the Public Health Service (PHS R01 NS 20314) and the National Science Foundation (BNS 8812728) to Massaro, a grant from the National Science Foundation (SES 90 23945) to Friedman, and a grant from the National Science Foundation (SBR-9310347) to Friedman and Massaro. We thank Bill Estes and Josh Hurwitz for making their data available to us, and Thomas Wallsten, Jerome Busemeyer, and two anonymous reviewers for editorial guidance and thoughtful comments on earlier versions of this paper. We dedicate this paper to the memory of N. John Castellan, Jr., a pioneer in the field of probabilistic learning.

0022-2496/95 $6.00

A BRIEF HISTORY

The study of learning has a long and multifaceted history within psychology. Experiments on learning in animals at the turn of the century were guided by association theory. A response to an environmental event that led to some reward would be more likely to occur at some future time. Thorndike, in his famous puzzle box experiments with cats, formulated a theory of connectionism in which one appropriate response became connected with another and erroneous responses were stamped out. William James anticipated the development of neural networks to describe how to behave appropriately given some environmental event.

The task we investigate here has roots in the 1950s probability learning experiments, in which human subjects predicted which of two events would occur on each of a series of trials. For example, the two events might be the left light with probability 0.8 and the right light with probability 0.2, independently on each trial. This paradigm placed the human subject in a simple nondeterministic situation. An extension, called discriminative probability learning, involved two test stimuli with independent reinforcement schedules (Massaro, 1989; Myers & Cruse, 1968).

To maximize the number of correct responses, the subject should *always* chose the more likely event given the stimulus, e.g., the left light. The persistent finding in this research was that subjects tended to probability match, e.g., to chose the left light with probability 0.8 (Estes, 1959). In a significant number of experiments, however, subjects tend to overshoot, i.e., to exceed probability matching (Massaro, 1969; Myers & Cruse, 1968).

Castellan (1974) studied multiple cue probability learning with up to four separate cues. In this and related studies he examined the effect of various sorts of feedback, especially in the form of a summary statistic (ϕ) for the normative utilization and/or the subject's actual utilization of each cue. The feedback typically brought asymptotic performance only modestly closer to maximizing the number of correct responses, the strongest effects appearing in experienced subjects receiving both actual and normative utilization. Substantial individual differences were found and it is not clear that all subjects understood this added information. Castellan (1977) reviews the large body of previous relevant literature.

Gluck and Bower (1988) revived the contemporary study of multiple cue probability learning in a medical diagnosis task. Gluck and Bower (1988) found that an adaptive network model gave a good description of subjects' asymptotic performance. In addition, they extended the standard learning task to include single-symptom test trials in which a single symptom was presented and subjects had to estimate the percentage of patients that would be expected to be suffering from a given disease. The results on special trials indicated that subjects tend to overestimate the probability of the rare disease, a form of base-rate neglect. Estes *et al.* (1989) extended this study to include tests of models' predictions of trial-by-trial learning. They found evidence for a particular network model over an exemplar model.

CONCEPTUAL FRAMEWORK AND MODELS

We consider a slight generalization of the Gluck and Bower (1988) task. A subject is presented in each trial with stimulus vector $\mathbf{s} = s_1 s_2 \cdots s_n$, where, for $i = 1, ..., n$, $s_i = 1$ or 0 indicates one value (1) or the alternative value (0) of a medical symptom i in a particular patient. For example, $s_1 = 1$ might indicate a sore throat and $s_3 = 0$ might indicate the absence of dizzy spells. The subject must respond by stating whether ($d = 1$) or not ($d = 0$) she believes that this patient has the target disease or, equivalently, which of two diseases a patient has. Then the subject is told the actual value of d. Our task uses $n = 4$ conditionally independent symptoms and a random trial sequence. The classical probability learning experiments can be regarded as degenerate medical diagnosis tasks with $n = 0$ symptoms; i.e., subjects must learn only a fixed numerical probability (Estes, 1959). The probabilistic discrimination experiments can be regarded as simple medical diagnosis tasks with $n = 1$ symptom (Massaro, 1969). In this task, subjects must learn only two independent probabilities to the two levels (values) of a single symptom.

The medical diagnosis task is very simple in that stimuli $\mathbf{s}$ and responses d are both composed of binary variables, coded here as 0 or 1. But with one or more symptoms presented the task is nontrivial for two reasons. First, all trials are training trials in that the subject always receives

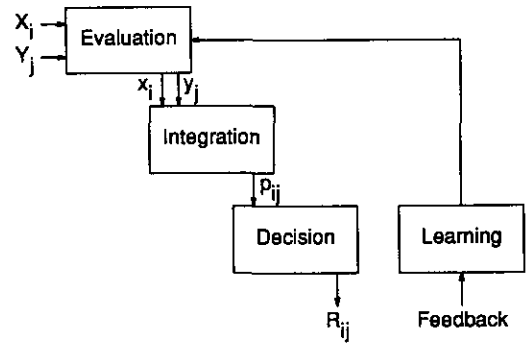

FIG. 1. Schematic representation of the three stages of information processing. The sources of information are represented by uppercase letters. The psychological representations are given by lowercase letters. The evaluation process transforms the symptom values into psychological values (indicated by lowercase letters). The integration process takes the output of the evaluation process and combines or integrates the psychological values to give an overall value for each of the relevant alternatives. The decision operation maps these values into some response, such as a discrete decision or a rating.

feedback (the actual value of d) and the subject has no initial knowledge of the relation between d and the given s. Second, the relationship between the feature values and the feedback is stochastic, so the subject sometimes will be misled. For example, the subject may pick a rather unlikely disease and turn out to be correct on that trial, or pick the highly likely disease and turn out to be incorrect.

The main goal of this paper is to find models that adequately describe subjects' behavior as they learn to perform the medical diagnosis task. It will be useful to describe the models within an analytic framework used by Massaro and Friedman (1990). The medical diagnosis task involves multiple sources of information, and these sources must be processed and mapped into a decision. As illustrated in Fig. 1, the framework assumes three stages of information processing intervening between presentation of the symptoms and the response: evaluation, integration, and decision. Here we add a fourth stage, learning or updating in response to feedback.

Three of the models (Bayes, FLMP, and CMP) share a number of important properties and also differ in specific ways. All three models assume that each of the symptoms is intially processed or evaluated independently of the other symptoms. Furthermore, continuous information is made available by the evaluation and integration processes. In addition, all of the models assume a relative goodness rule—RGR (Massaro & Friedman, 1990) or probability matching at decision. The models differ from one another in terms of the interpretations they provide to the parameters or what Massaro and Friedman (1990) call currency. The Bayes model and the FLMP assume mathematically equivalent rules for the integration stage. However, their updating of the parameter values during learning differs. The Bayes model differs from the CMP in terms of both integration and updating during learning. The FLMP differs from the CMP only in terms of the integration process. Even with this difference, these two models make equivalent predictions for tasks with two mutually exclusive response alternatives (but not for tasks with three or more alternatives). The exemplar model differs from these three models by storing each symptom configuration and the associated disease. The ALCOVE model is more complex by combining aspects of the exemplar model and connectionism. We now develop the models to be tested.

Bayes

The Bayesian model is normative in that it assumes no loss of information or biases introduced by irrelevant information. In this case, each learning experience is processed fully and is used completely in updating memory. Furthermore, it is assumed that the subject begins the experiment *tabula rasa*. On the first trial, the subject necessarily responds randomly, and only observations made during the actual experiment have any influence on performance. In our strict interpretation of the Bayesian model, we do not allow any information about base rates or conditional frequencies other than those actually observed in the experiment. That is, we assume that subjects process information as if they were Bayesian statisticians with initially diffuse priors.

Bayes theorem can be expressed either in terms of prior and posterior probabilities or, alternatively, in terms of prior and posterior odds. For p and $\bar{p}$ representing the (prior or posterior) probabilities of $d=1$ and $d=0$, we can write the odds ratio as $r=p/\bar{p}=p/(1-p)$, and can recover the probabilities from the odds ratio by $p=r/(1+r)$ and $\bar{p}=1/(1+r)$. The odds ratios lead to more convenient expressions, so the derivation starts with Bayes theorem in odds form:

$$\frac{P[d=1\,|\,s]}{P[d=0\,|\,s]}=\frac{P[s\,|\,d=1]}{P[s\,|\,d=0]}\cdot\frac{P[d=1]}{P[d=0]}. \tag{1}$$

The verbal statement of (1) is that the posterior odds are equal to the likelihood ratio times the prior odds.

Next, use conditional independence to expand the likelihoods $P(s\,|\,d)=P(s_1\,|\,d)\,P(s_2\,|\,d)\cdots P(s_n\,|\,d)$. We reduce clutter by writing the elementary likelihoods as $p_i=P(s_i=1\,|\,d=1)$ and $q_i=P(s_i=1\,|\,d=0)$. Note that in general p_i and q_i can vary independently, but that $P(s_i=0\,|\,d=1)=1-p_i$ and $P(s_i=0\,|\,d=0)=1-q_i$. Since $s_i=0$ or 1, we note that $p_i^{s_i}(1-p_i)^{1-s_i}$ is p_i if $s_i=1$ and is $1-p_i$ if $s_i=0$. Using this and a similar expression for the q_i's, we expand the likelihood ratio in (1) to arrive at

$$\frac{P[d=1\,|\,s]}{P[d=0\,|\,s]}=\left[\left(\frac{p_1}{q_1}\right)^{s_1}\left(\frac{1-p_1}{1-q_1}\right)^{1-s_1}\cdots \times\left(\frac{p_n}{q_n}\right)^{s_n}\left(\frac{1-p_n}{1-q_n}\right)^{1-s_n}\right]\frac{P[d=1]}{P[d=0]}. \tag{2}$$

Equation (2) expresses the posterior odds in terms of the $2n$ numerical values p_i and q_i, $i=1,\ldots,n$, plus another numerical value for the prior odds. Let $Y(s)$ represent the posterior odds, the left-hand side of (1) or (2), and let $b_o=P[d=1]/P[d=0]$ represent the prior odds (or "base odds"). Grouping the fractions we get the more compact expression

$$Y(s)=b_o\prod_{i=1}^{n}\left(\frac{p_i}{q_i}\right)^{s_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-s_i}. \tag{3}$$

In terms of the information processing framework, the Bayes model evaluates each symptom i as a feature value p_i or q_i and integrates them using the multiplicative formula (3). For the decision stage, we assume that subjects choose

their responses randomly according to the computed posterior odds. That is, we assume the relative goodness rule or probability matching (Massaro & Friedman, 1990). This decision rule is not normative if subjects attempt to maximize the number of correct responses (although it may be normative if subjects have other goals). We assume RGR at the decision stage in all other models and do so here to maintain comparability. The "currency" or interpretation of the variables is subjective probability.

How might subjects come to learn the parameter values p_i and q_i and b_o? A Bayesian statistician with no prior knowledge except that the symptoms are conditionally independent and the trials are exchangeable (randomly sequenced) would estimate p_i simply by counting the number of times $s_i = 1$ and $d = 1$ relative to the number of times $d = 1$. Analogous count ratios would be used to estimate the q_i's. The value b_o would be estimated by counting the number of times $d = 1$ and dividing by the number of times $d = 0$. Denoting these ratios as p_{it}, q_{it}, and b_{ot} and substituting them into Eq. (3), we have the Bayesian model's prediction $Y_t(s)$ of the odds that a subject would choose the first disease if presented with the symptom configuration s on trial t.

Fuzzy Logical Model of Perception (FLMP)

The fuzzy logical model of perception (FLMP) formalizes developments in fuzzy logic, pattern recognition, and choice theory to provide a systematic account of perceptual judgment and decision making (Massaro, 1987). Its currency is a truth value that can range from 0 or completely false to 1 or completely true. For example, a 0.3 truth value for the proposition "a whale is a fish" means that the proposition is true to degree 0.3; i.e., there is a moderate degree of similarity between whales and fish. It is psychologically different from a probability. An objective 0.3 probability that "a whale is a fish" would mean that 3 out of 10 whales are fish. A subjective 0.3 probability would mean that that a bet (that a whale is a fish) with 7 to 3 odds is fair. Neither of these probabilistic interpretations works because a whale is never a fish.

According to the FLMP, feature evaluation in the medical diagnosis task produces a truth value (or "feature value") denoted f_i for the statement "the disease is $d = 1$" when $s_i = 1$, and produces a value g_i for the same statement when $s_i = 0$. The two diseases are mutually exclusive and exhaustive, so one uses the fuzzy logical negation operation to conclude that the truth values for the statement "the disease is $d = 0$" are $1 - f_i$ and $1 - g_i$ when $s_i = 1$ and $= 0$, respectively. In the absence of symptoms the statement "$d = 1$" has some truth value denoted b. As explained in Massaro and Friedman (1990), the integration stage of information processing in the FLMP combines the feature values to produce integrated truth values, using the same multiplicative formula as in Bayes. That is, even though the feature values are psychologically different from elementary conditional probabilities, the integration stage treats them in an analogous fashion. Specifically, the result of FLMP integration in the medical diagnosis task is a truth value for $d = 1$ of

$$\frac{b \prod_{i=1}^{n} f_i^{s_i} g_i^{1-s_i}}{b \prod_{i=1}^{n} f_i^{s_i} g_i^{1-s_i} + (1-b) \prod_{i=1}^{n} (1-f_i)^{s_i} (1-g_i)^{1-s_i}}$$

after observing the symptom configuration $s = s_1 \cdots s_n$. In ratio form, the expression becomes

$$Y(s) = b_o u_1^{s_1} v_1^{(1-s_1)} \cdots u_n^{s_n} v_n^{(1-s_n)}, \tag{4}$$

where b_o is $b/(1-b)$, $u_i = f_i/(1-f_i)$, and $v_i = g_i/(1-g_i)$.

The decision stage again is RGR—the odds with which subjects choose the diseases are given by Eq. (4). It is clear that Eqs. (3) and (4) are equivalent. In our implementation of the models, however, we use objective frequencies for the Bayesian model and subjective truth values for the FLMP.

The two models also differ from one another with respect to the learning or the updating of the feature values from trial $(t-1)$ to trial (t). Learning in the FLMP is described by the general rule

$$\begin{aligned} u_t &= u_{t-1} + \lambda e s_i, \\ v_t &= v_{t-1} + \lambda e (1 - s_i), \end{aligned} \tag{5}$$

where u and v denote the features being learned, λ denotes the learning rate, and $e = d - p(d \mid s)$ denotes the perceived error, given the current feedback d and truth value $p(d \mid s)$ of that disease assessed using the current feature values for the current symptoms s. An analogous formula can be written for b_o but it seems to require a much different learning rate. As explained in the Appendix, the ratio b_o can be absorbed into the other ratios u_i and v_i, and so we estimate Eq. (4) without the b_o parameter. Thus, in a four-symptom task, there are four f_i and four g_i for a total of eight feature values. Note that the g_i but not the f_i are updated when $s_i = 0$, and the reverse is true when $s_i = 1$. All feature values f and g are assumed to start at 0.5 (completely ambiguous) in the FLMP. We call this model FLMP8.

Equation (5) reveals several characteristics of the updating rule (or learning algorithm) for the FLMP. First, learning occurs only to the extent that there is nonzero perceived error. Second, all features (symptoms) are learned at the same rate λ. Third, the learning rate λ is constant throughout the course of learning. We use this same updating rule in the test of all of the models except Bayes to keep them as comparable as possible.

Connectionist

Much of the revival of interest in probabilistic learning comes from the renewed excitement over neural networks or

connectionist models. Connectionist models map input to output via activation along a number of simple processing units organized into separate layers. Here we consider only the simplest sort of connectionist models with two layers, an input layer and an output layer. The stimulus input is mapped to a set of input units that are connected to a set of output units that are, in turn, connected to responses. Activation is propagated from input to output and modulated by the weights on the connections between the input and output units. These weights are dynamically modified during learning. Information about the symptoms and diseases is contained in the weights connecting the input and output units.

We focus on a simple connectionist model called CMP_{n+1} and illustrated in Fig. 2. It has one output node and n input nodes, one for each symptom plus a bias (or base rate) node constantly at unit activation. Feature evaluation consists of the activation $w_i s_i$ produced by each node in the input layer ($w_0 1 = w_0$ for the base rate node). Given four symptoms in our task, there are four w_i and a w_0 parameter for a total of 5. We call this model CMP5. Information integration consists of summing the activation arriving at output node $y = w_0 + \sum_{i=1}^{4} w_i s_i$ and using the standard sigmoid function $S(y) = 1/(1+e^{-y})$ to rescale the output to the range $[0, 1]$. As usual, the decision stage is RGR—disease 1 is chosen with probability $S(y)$ and the alternative disease with probability $1 - S(y)$. In the current task (or any other task with two mutually exclusive response alternatives), it has been shown that this model is asymptotically equivalent to the FLMP (Massaro & Friedman, 1990).

The learning algorithm for connectionist models is standard; weights are updated after each leaning trial according to the perceived error. For the medical diagnosis task the learning algorithm is

$$w_{it} = w_{it-1} + \lambda(d - S(y|s))\, s_{it-1}. \tag{6}$$

(For w_0 use the convention $s_0 = 1$ on every trial.) Clearly this is the same learning algorithm as used in the FLMP model and so the models are formally equivalent for our task.

Gluck and Bower (1988) investigated an asymptotic version of a four-input node (plus one output node) model in a four-symptom medical diagnosis task. Estes *et al.* (1989) tested two other versions of the connectionist model for the standard four-symptom medical diagnosis task. In the Appendix we show that these (and some other two-layer connectionist models) are very closely related to our five-parameter CMP5. We also show there that, despite their quite different underlying psychology, the connectionist models have a very close formal relation to the Bayesian model.

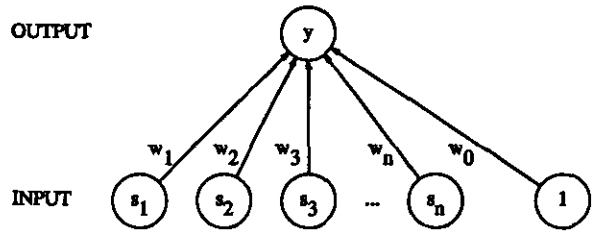

FIG. 2. Illustration of CMPn + 1 connectionist model.

Exemplar

Exemplar models are fundamentally different from models based on summary descriptions. Models based on summary descriptions potentially are more efficient because these models select the relevant information about each feature, categorize the input based on the summary description, and update it based on feedback. Exemplar models store the complete symptom configuration and the associated disease. At test, the exemplars in memory are used to categorize the input. Exemplar models can assume attention weights or allow a nonzero similarity among different symptom configurations. Estes *et al.* (1989) (and previously Medin, Altom, Edelson, & Freko, 1982) applied an exemplar model to the medical diagnosis task. Here we will test a simple version of the Estes *et al.* (1989) model; its detailed specification and relation to other exemplar models are presented in the Appendix.

In the exemplar model we test, each of the symptom configurations $j = 1, ..., 2^n$ is considered quite distinct from the others. Each configuation j has a weight w_j that reflects the fraction of stored exemplars associated with the target disease ($d = 1$). At evaluation stage, the weight w_j is transmitted directly to the decision stage if j corresponds to the actual stimulus configuration; the weights corresponding to other configurations are not transmitted. The integration stage is bypassed because there is only one nonzero output from feature evaluation. The output, transformed by the usual sigmoid function $S(w_j) = 1/(1+e^{-w_j})$, is the predicted probability of responding with that category.

We apply the standard learning algorithm to the weights, exactly one of which is updated after each trial. The Appendix shows that this version of the exemplar model is an appropriate simplification of the model used by Estes *et al.* (1989).

ALCOVE

Our final candidate for explaining behavior in the medical diagnosis task is the ALCOVE model (Kruschke, 1992)—a hybrid of connectionism and spatially represented exemplars. The Alcove model extends the exemplar model presented above by including a parameter for similarity scaling, so all exemplar representations have a chance to influence performance. We use the version of ALCOVE adapted to the medical diagnosis task by Nosofsky *et al.*

(1992). This version does not include the mechanism for attention-strength learning.

The ALCOVE model has three layers. The input layer has four nodes, one for each of the four symptoms. Feature evaluation consists of the activation a_i^{in} of these nodes. Information integration is more complex. The activation a_j^{hid} of each of the 2^4 exemplar hidden units $j = 1, ..., 16$ arises from feature evaluation according to the formula

$$a_j^{hid} = \exp\left[-c \sum_i |h_{ij} - a_i^{in}| \right], \qquad j = 1, ..., m, \qquad (7)$$

where c is a positive constant and the h_{ij} represent the n-dimensional locations of the exemplar units. The effect of the $|h_{ij} - a_i^{in}|$ term is to activate all hidden units according to their similarities, which are then summed. Information integration is completed by summing the activations a_j^{hid} from the hidden units, modulated by the learned weights w_{kj} at each output node $k = 0, 1$ according to

$$a_k^{out} = \sum_j w_{kj} a_j^{hid}. \qquad (8)$$

At the decision stage these activations are mapped into response probabilities by an RGR on the exponentially transformed activations

$$P(k) = \exp(\phi a_k^{out}) \Big/ \sum_k \exp(\phi a_k^{out})$$

using a free parameter ϕ. It turns out that in the medical diagnosis task the second output node is redundant; given the standard weight initialization, we get exactly the same fits with one output node. The ϕ parameter also is redundant in that it can be constrained to be 1.0 without affecting the fit to the data. Thus we can use the standard sigmoid function so the last part of information integration (and the decision stage) of ALCOVE is exactly the same as for the exemplar model.

The main difference between ALCOVE and our exemplar model is the similarity parameter c. When c is larger than 1.0, the influence of the hidden unit j is disproportionately large and the influence of more distant units is disproportionately small. When c is sufficiently large, the activation of all hidden units is effectively zero except for the unit j that corresponds to the stimulus vector s. Then ALCOVE in effect reduces to our basic exemplar model.

Learning is confined to the weights w_{kj} from the hidden layer to the output layer. The formula again is given by Eq. (6). The similarity parameter c and the learning rate λ are estimated to give the best fit to the data.

PRESENT STUDY

The present study is a partial replication and extension of the Gluck and Bower (1988) and Estes *et al.* (1989) studies. We modified the typical symptom probabilities used in these tasks because they produced only nine unique posteriori probabilities for the 16 symptom configurations. With our new symptom probabilities, each symptom configuration has a unique posteriori probability of occurrence. We did not include interspersed test trials on just one of the symptoms (as was done in those studies). The reason is that we were concerned whether the subjects would interpret this new type of trial appropriately. It is possible that subjects interpret the new type of trial with a single symptom as coming from a different population than the four-symptom configurations. Accordingly, these test trials might be an unfair test of the normative and descriptive models. We did include, however, a condition with symmetric base rates as well as the standard one with asymmetric base rates. These two conditions allow performance to be tested at a much larger range of posteriori probabilities. With asymmetric base rates, most of the symptom configurations have relatively extreme posteriori probabilities for the disease with the highest base rate. With symmetric base rates, the posteriori probabilities are more symmetric and less extreme. These extensions allow a broader empirical base for comparing the various models. They also allow us to test for stability of parameter estimates across the two base rates.

METHOD

Subjects

Forty-two members of the university community participated in the experiment for about 1.5 h each. Their participation fulfilled a course option or they were paid $5.56 per hour.

Apparatus

All experimental events were controlled by a DEC PDP-11/34A computer. Four sound attenuated subject rooms were used, each illuminated by two 60-W incandescent bulbs in a frosted glass ceiling fixture. Each room contained a chair facing a table holding the visual display, a TVI950 terminal.

Procedure

There were two sessions of 240 trials, each taking about 30 min with a 5-min break in between. Subjects were told that this was a study of how people learn from experience and make decisions. The medical diagnosis task, the diseases, and the symptoms were described. Subjects were

instructed that there were two possible diseases and four symptoms. Each of the symptoms could be negative ("−") or positive ("+"). Subjects were told to evaluate the symptoms and decide if the patient had one disease or the other. The feedback was described and it was stressed that the order of patients was strictly random with no particular sequence of diseases that could be guessed.

Stimuli

Each trial began with a vertical table of the four symptom names: Circulation, Temperature, Pain, Skin. The values of these symptoms were indicated by placing a "+" or a "−" next to each of the symptom names. Each subject then responded by pushing one of two buttons labeled by the disease names. The response interval was not limited. After a subject responded, feedback was given by displaying the appropriate names of the diseases next to the "ACTUAL ILLNESS" and "YOUR DIAGNOSIS" entries. Given that up to four subjects could be tested simultaneously with yoked displays, the symptom and feedback values were not cleared from the display until 2 s after the last subject responded. The next trial began 3 s later.

The stimuli were generated following the method of Estes *et al.* (1989). In each trial the subject was shown a configuration of four symptoms exhibited by a particular patient, and was asked to provide a diagnosis. The two possible diseases were called ROMELLA and BIRNOMA and the four symptoms were called Circulation, Temperature, Pain, and Skin. For each dimension, the symptom could be positive (+) or negative (−). For analytical purposes we coded the symptoms as s_1 to s_4 with values 0 and 1, and coded the diseases as category A (or $d=1$) and B (or $d=0$).

Table 1 gives the likelihood (p_i, q_i) of each symptom for disease categories A ($d=1$) and B ($d=0$). We used two base rates for the occurrence of category A: 0.5 (referred to as symmetric) and 0.25 (referred to as asymmetric). The stimuli were generated in two blocks of 240 trials each. Within each block the frequencies of each category and each feature were constrained to match the probabilities in Table 1. It should be noted that the higher order combination frequencies were not constrained, except that random sequences were chosen for each block of 240 which ensured that, for both base rates, all 32 possible stimulus events (four dimensions with two levels per dimension for both categories) occurred at least once. All subjects tested at a given base rate received the same sequence of test stimuli.

TABLE 1

Probability of Sympton Occurrence for Categories A and B for the Original Estes *et al.* (1989) and the Current Study

Data set	Estes *et al.*		Current	
Category	A	B	A	B
Symptom				
1	0.6	0.2	0.6	0.25
2	0.4	0.3	0.4	0.35
3	0.3	0.4	0.3	0.45
4	0.2	0.6	0.2	0.65

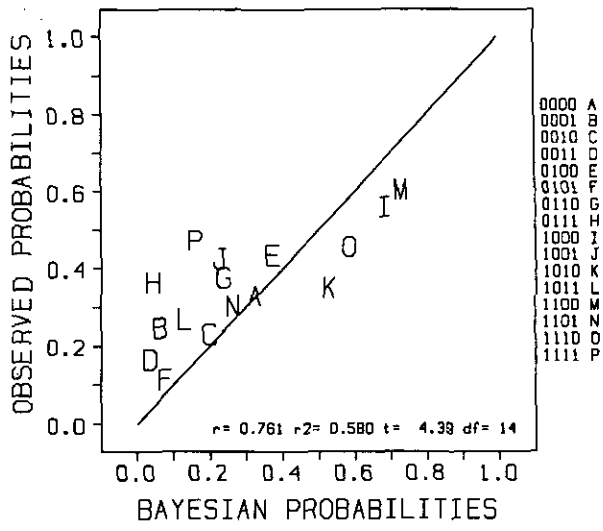

FIG. 3. Observed probabilities of category A responses to each of the 16 stimulus combinations as a function of their a posteriori Bayesian probabilities for the asymmetric condition in the first block of 240 trials.

RESULTS

The first question is to determine if subjects learned to respond differentially to the different symptom configuration To answer this question, we divided the 480 trials in

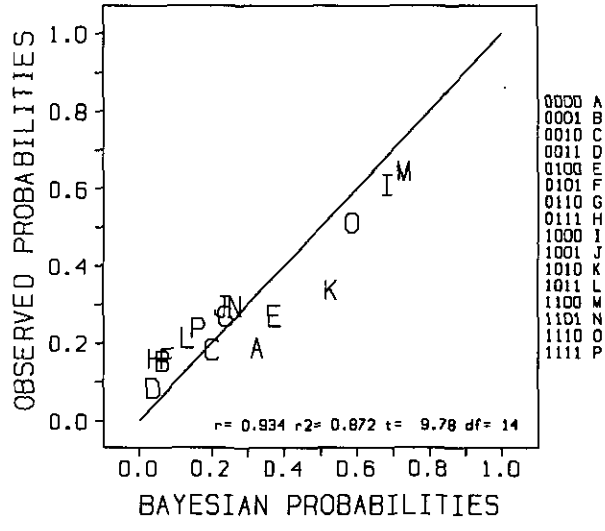

FIG. 4. Observed probabilities of category A responses to each of the 16 stimulus combinations as a function of their a posteriori Bayesian probabilities for the asymmetric condition in the second block of 240 trials.

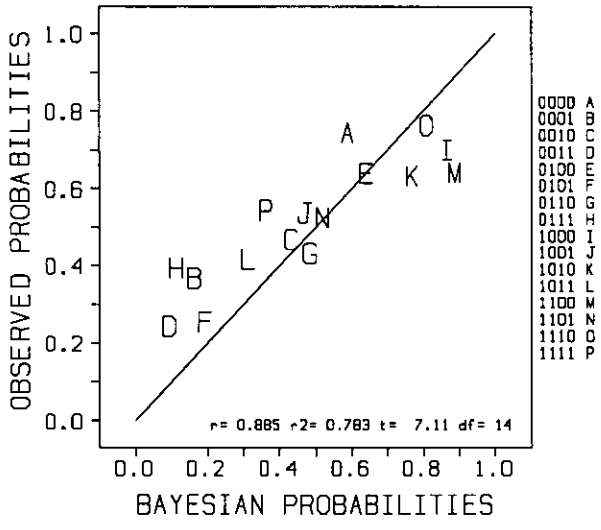

FIG. 5. Observed probabilities of category A responses to each of the 16 stimulus combinations as a function of their a posteriori Bayesian probabilities for the symmetric condition in the first block of 240 trials.

each data set into two blocks of 240 consecutive trials. For each block and for each symptom configuration we correlated the true probability of disease A (given the symptom configuration) with the observed frequency with which the 24 subjects responded A. The true probability of a disease is equivalent to the a posteriori Bayesian probability. If subjects learn the probabilities conditioned on the symptom configurations (and if they probability match), then the correlation will increase across the two trial blocks.

Figures 3 and 4 plot the observed response probabilities against the a posteriori Bayesian probabilities in the asymmetric condition for each of the 16 symptom configurations for the two blocks of trials. As can be seen in the figures,

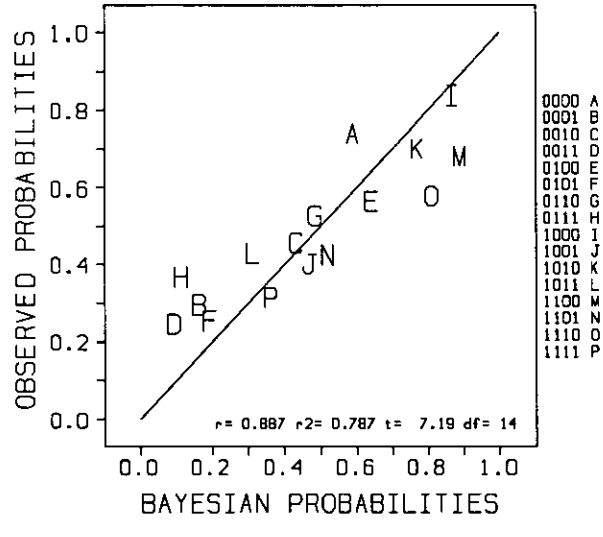

FIG. 6. Observed probabilities of category A responses to each of the 16 stimulus combinations as a function of their a posteriori Bayesian probabilities for the symmetric condition in the second block of 240 trials.

subjects learned to approximate the conditional probabilities. The correlation increased from 0.761 in the first block to 0.934 in the second block. Figures 5 and 6 plot the observed response probabilities against the a posteriori Bayesian probabilities in the symmetric condition. Although learning did not occur across these two blocks in the symmetric condition, it appears that learning was completed earlier in the experiment. When the trials were partitioned in blocks of 160 trials each, the correlation increased from 0.703 in the first block to 0.912 in the second block. Performance became somewhat more variable in the third block with the correlations dropping slightly to 0.860. We conclude that subjects learned to respond differentially to the different symptom configurations, so it makes sense to compare the predictions of the various models against the asymptotic and learning results.

The asymptotic and trial-by-trial predictions of each model were fit to the data using a program written by the third author. The observed data to be fit are the responses pooled across subjects, i.e., the fraction of subjects choosing disease A ($d=1$). The main reason is that the theoretical models predict some probability of an A response (a point on the 0–1 continuum) while the task allows each subject only the binary (endpoint) response 0 = B or 1 = A. (Tests of the models against individual data would be possible if the task permitted a continuous response such as estimation of probability of a given disease category.) The asymptotic data consist of 16 observations: subjects' responses averaged across the 21 subjects and the repeated occurrences in an appropriate block of trials of each of the 16 symptom configurations. The trial-by-trial data consist of 480 observations: responses averaged across the 21 subjects for each of the 480 individual trials. Each model is represented as an algorithm for computing the sum of squared deviations between the observed and predicted data as a function of a set of parameters. By iteratively adjusting the parameters of the model, the program minimizes the sum of squared deviations (or equivalently, the root mean squared deviation, RMSD) between the observed and predicted points. Alternatively, the program maximizes the likelihood (or log likelihood) of the data. Thus, the program finds a set of fitted parameter values which when put in the model come closest to predicting the observed data.

To provide a measure of the absolute goodness-of-fit of a model, we computed a *benchmark* RMSD, the RMSD that would be expected if the model were correct. For all of the models, the output of the integration stage is mapped into a response probability via a relative goodness rule (RGR), independently for each subject on each trial. Thus, even if a model were correctly specified and the fitted free parameters (e.g., connection weights) were exactly correct, there would be a binomial sampling error for the response probabilities averaged across subjects.

The benchmark RMSD is defined as the binomial RMSD calculated from the actual data by the formula $n^{-1}pq$. For example, suppose that 16 of 21 subjects chose disease A ($d=1$) on trial 123. The binomial variance on this trial is $21^{-1}(16/21)(1-16/21)\approx 0.00864$, and the standard deviation (the RMSD for the single trial) is about 0.093. The benchmark RMSD for a set of trials is the square root of the mean binomial variance (the binomial variance for each trial averaged across the trials).

The RMSD for each fitted model can then be compared to the benchmark RMSD. When the benchmark RMSD is reasonably small and closely approximated by the actual RMSD we can conclude that the model does an absolutely good job of explaining the data. When all models have actual RMSD greatly in excess of their benchmarks, we can conclude that even the model with the lowest RMSD does not adequately capture subjects' average behavior. In this case we should seek new models and/or more refined data.

ASYMPTOTIC FITS

To assess asymptotic performance, three of the models were fit to the observed probabilities of responding disease A ($d=1$) to each of the 16 symptom configurations averaged over the last block of 240 trials in the two current data sets (symmetric and asymmetric conditions), and averaged over the last 120 trials in the two shorter data sets of Estes *et al.* (1989). The exemplar and ALCOVE models were not tested because the number of free parameters for these models would equal or exceed 16, the number of independent data points.

Table 2 presents the models and the results of their fit to the asymptotic data. The asymptotic Bayes (normative) model predicts that the probability of responding disease A to each symptom configuration should be equal to the objectively correct conditional probability of disease A given that symptom configuration, shown for each data set in Figs. 3–6. This model has no free parameters (NP = 0). Its RMSD values for the four data sets range from 0.136 to 0.091, about 3 and 5 times larger than the benchmark RMSDs. We conclude that the Bayesian model gives a poor description of the asymptotic results.

The FLMP is equivalent to the normative Bayesian model but allows the subjective feature support values (and prior odds) to differ from their objective probabilities. The eight free parameters in the FLMP8 reduce the RMSDs considerably, as can be seen in Table 2. As expected, the five free parameter CMP5 gave the same asymptotic predictions as the FLMP. As mentioned in the derivation of the models, it has been shown that these models are asymptotically equivalent for two mutually exclusive response alternatives (Massaro & Friedman, 1990). However, the RMSDs still were between 2 and 3 times larger than the benchmarks, so the asymptotic performance of these descriptive models is still well below an adequate fit.

TABLE 2

The Number of Free Parameters (NP) and the RMSD Values for the Models Fit Asymptotically

		Symmetric	Asymmetric		
Model	NP	Current	Current	Group E	Group C
Bayes	0	0.132	0.091	0.103	0.136
FLMP	8	0.065	0.044	0.070	0.086
CMP	5	0.065	0.044	0.070	0.086
Benchmark		0.026	0.024	0.030	0.029

Note. The fit is to response average in the last 240 trials of the symmetric and asymmetric conditions of the current study, and in the last 120 trials of the estimation (Group E) and categorization (Group C) conditions of Estes *et al.* (1989).

TRIAL-BY-TRIAL LEARNING

Trial-by-trial data can be expected to produce larger RMSDs than the asymptotic data for at least three reasons. First, an observation in the trial-by-trial data consists of a smaller sample of subjects' choices than an observation in the asymptotic data. Second, it is harder to capture a changing process than a steady state. Third, it is harder, other things being equal, to fit 480 (or 240) observations than to fit 16 observations. The first reason leads to larger benchmark RMSDs, e.g., about $\sqrt{120/16}\approx 2.7$ times larger for the last 120 trials (observations) in the Estes data than for an observation in the asymptotic data (16 observations).

Table 3 summarizes the trial-by-trial fits and, as expected, the RMSDs are all larger than for the asymptotic fits. The Bayesian model produces RMSDs approaching 0.20, indicating fairly sizable and frequent errors in predicting subjects' average choice. Of course, this model has no free parameters. The FLMP8 and the CMP5 learning models

TABLE 3

The Number of Free Parameters (NP) and the RMSD Values for the Models Fit Trial by Trial

		Symmetric	Asymmetric		
Model	NP	Current	Current	Group E	Group C
Bayes	0	0.195	0.190	0.178	0.186
FLMP	1	0.167	0.159	0.168	0.181
CMP	1	0.166	0.157	0.165	0.177
Exemplar	1	0.150	0.153	0.145	0.157
ALCOVE	2	0.150	0.150	0.133	0.143
Benchmark		0.095	0.087	0.078	0.074

Note. The fit is to trial-by-trial results of the symmetric and asymmetric conditions of the current study, and in the last 120 trials of the estimation (Group E) and categorization (Group C) conditions of Estes *et al.* (1989).

have a single free parameter, the learning rate λ that adjusts the internal values (eight truth values or five connection weights, respectively). The free parameter allows these descriptive models to noticeably improve performance over the Bayes model. The CMP5 consistently did a shade better than the FLMP8, suggesting that the five internal value scheme is a bit more efficient than the eight-value scheme. As shown in the Appendix, the five-parameter scheme has a natural FLMP interpretation, and one could easily construct a network model mathematically equivalent to the eight-value FLMP. Given their mathematical equivalence, the slight edge of CMP5 over FLMP8 should not be interpreted as proving the superiority of an underlying psychological theory.

The ALCOVE learning model gave the best fit, using the free parameters λ and c. However, the exemplar model uses only one free parameter, fits the current data almost as well as ALCOVE, and does relatively well on the Estes data. Although the descriptive models improve on the fit given by the Bayesian model, there is still considerable room for improvement. The RMSDs for the descriptive values, shown in Table 3, vary between 2 and 1.5 times the benchmark RMSDs.

Table 3 is a very brief summary of our empirical work. We constructed a variety of connectionist and other models and looked at some alternative learning algorithms. None of these variants outperformed the basic models reported in the table. We also found basic model fits that maximized likelihood instead of minimizing RMSD, but we found no substantial change in the parameter estimates.

Maximum likelihood estimation does permit additional tests. The Appendix shows that the ALCOVE model reduces to the basic exemplar model as the similarity parameter $c \to \infty$. The estimates of c typically are fairly large and the RMSDs for ALCOVE are sometimes only a bit less than for the basic exemplar model. Could it be that the exemplar model in fact is correct? The question can be answered using maximum likelihood techniques. Under the null hypothesis that the constrained model is correct, it can be shown that twice the difference between the maximized log likelihood of the constrained model and the maximized log likelihood of the unconstrained model has the χ^2 distribution with k degrees of freedom, where k is the number of constraints (e.g., Cramer, 1986). For the question at hand, $k = 1$ and we obtain the results shown in the first three lines of Table 4. Thus we can confidently say that the ALCOVE model explains the trial-by-trial mean choices better than the basic exemplar model.

Is our benchmark RMSD too demanding? A maximum likelihood approach is to form a saturated model that will fit the data perfectly because it has as many parameters as there are data points. If one of our models is exactly correct, then the difference in log likelihoods between it and the saturated model with have a χ^2 distribution as above. The ALCOVE fits are best and therefore have the best chance of passing the test. But the results shown at the bottom of Table 4 show we can strongly reject the null hypothesis. We conclude that the ALCOVE model (and also the other models) does not provide a complete explanation of the trial-by-trial data.

TABLE 4

Chi Squared Tests of the ALCOVE versus the Exemplar Model and the Saturated Model versus the ALCOVE model for the Symmetric and Asymmetric Conditions

	Symmetric	Asymmetric		
Model	Current	Current	Group E	Group C
LL(ALCOVE)–LL(Exemplar)	17.17	107.22	47.29	77.37
Chi-squared(1)	34.3	214.4	94.6	154.7
Significance, $p <$	0.001	0.001	0.001	0.001
LL(saturated)–LL(ALCOVE)	570.78	638.39	267.09	343.18
Chi-squared(478)	1141.6	1276.8		
Chi-squared(238)			534.2	686.4
Significance, $p <$	0.001	0.001	0.001	0.001

Note. Conditions cover the current study, the estimation (Group E), and the categorization (Group C) of Estes *et al.* (1989). The difference in log likelihoods (LL) is given, together with the number of degrees of freedom, and the significance level.

DISCUSSION

Our data point to three conclusions. First, subjects do learn in the medical diagnosis task. In the current (480 trial) data sets, subjects' average responses more closely match the objective conditional probabilities in later trials than in earlier trials, as illustrated in Figs. 3–6.

Second, all of the descriptive models did better than the normative (Bayesian) model in predicting subjects' asymptotic responses and trial by trial learning. The best performance was by the model with the most internal variables and free parameters, ALCOVE; but the basic exemplar model does almost as well with the same internal variables and only one free parameter.

Third, none of the models (neither the basic models reported here nor the variants discussed in the Appendix) provides an impressive fit to the data. The basic models produced errors 2 to 4 times larger than the benchmark in the asymptotic data and, more importantly, the learning models produced errors about 1.5 to 2 times larger than their benchmarks in the trial-by-trial data. Maximum likelihood techniques confirm the gap between the best fitting model (ALCOVE) and actual behavior. We conclude that there is considerable room for improvement.

These results provoke new questions. We began our work with the belief that we would be able to identify some learning model as the best of the current crop and would be

able to say why it was better than its rivals. Conclusions 2 and 3 above indicate why we hesitate to announce a winner. The ALCOVE model has the best fit but needs a lot of internal variables and free parameters to do so. More parsimonious models do almost as well on several data sets. Given the sizable gap remaining from the benchmark and given considerable instability of parameter estimates across data sets, we still regard it as an open question which learning models can best account for the data.

How might the learning models be improved in the future? In exploratory work, we find that the performance of the Bayesian model improves considerably when we relax the *tabula rasa* assumption of diffuse priors and allow a single free parameter to represent prior beliefs. This version of Bayes outperforms several of the behavioral models on some data sets and is much more parsimonious than any of them. On the other hand, when we use a learning algorithm similar to the Bayesian procedure (in particular, letting the learning rate λ decline as evidence accumulates or letting the error specification depend directly on the feature value as in Bayesian updating) we do not generally improve the fits of the behavioral models.

The models we considered differed considerably in their description of the three stages of information processing, but all of them emphasized learning the feature values, and we assumed that the information integration stage and the decision stages were fixed during the course of an experiment. Perhaps learning (behavioral change) takes place for the later two stages as well as for the evaluation stage. The old Markov learning models allow change at the decision stage, enjoy empirical success in some contexts (e.g., Massaro, 1969), and provide a natural way to incorporate a recency bias. Further theoretical work in this direction may be useful.

We believe that more refined experiments and data analysis may turn out to be at least as useful as more sophisticated models. In retrospect the experimental methodology used here and in previous studies has several remediable shortcomings.

First, there is a serious problem in interpreting the data unless we have a clearer idea of subjects' decision processes. The normative models and all of the descriptive models are applied to the data with the assumption that subjects will independently probability match or use the relative goodness decision rule (RGR). (See Massaro & Friedman, 1990, and Yellott, 1977, for another theoretical justification involving optimal deterministic decision-making based on noisy inputs.) It is important to observe that information processing can be optimal, even though the information transmitted by the subject is not veridical. The data would be easier to interpret if subjects had a clear and compelling goal at the decision stage.

Second, probably the most important shortcoming in current methodology is that the models are applied to aggregate data, assuming implicitly that all subjects obey the same learning model and that all have the same learning rate. We do not know of any evidence justifying these assumptions.

These two problems can be greatly reduced by changing the response mode of the experimental task. If subjects are allowed to give continuous responses (probability estimates, Wallsten, 1971) on each trial, and are capable of using an interval scale, then the RGR seems theoretically justified because it assumes that the subject has available the relative goodness of match of each alternative for a given symptom configuration. In this case, a subject's continuous response should be equal to the relative goodness of match of each alternative for a given symptom configuration. The aggregation problem can also be eliminated by elicited continuous (estimate) responses because these data can be fit trial-by-trial (or asymptotically) to the models for each subject separately.

Previous medical diagnosis-type studies have used estimation judgments or similar continuous responses only sporadically and have not tested models against the data. There has always been some controversy about whether estimation judgments can be considered to be an interval or linear scale of the representation or process of interest (Anderson, 1982). Although one cannot casually assume that subjects are using an interval scale, there are safeguards and tests that can be employed to ensure that responses are meaningful (e.g., Oden, 1978; Varey, Mellers, & Birnbaum, 1990). Estimation judgments, although less customary than categorical responses in medical diagnosis-type tasks, might be just as legitimate and potentially more informative.

APPENDIX: ANALYTICAL DETAILS

In this appendix we show that, despite their very different psychological origins, the behavioral models and the normative model have very close mathematical relationships. We also discuss alternative specifications of the behavioral models and the learning algorithms. Each model has a set of internal parameters (e.g., feature values) whose values are learned from experience. Models may differ asymptotically —i.e., because the parameter sets are essentially different— or they may differ only in that they use different learning algorithms to adjust the internal parameters. We begin with asymptotic comparisons, and discuss learning algorithms at the end.

Asymptotic Comparisons

Recall that the normative Bayesian model for n conditionally independent symptoms has the form

$$Y(s) = b_0 \prod_{i=1}^{n} \left(\frac{p_i}{q_i}\right)^{s_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-s_i},$$

where $Y(s)$ is the posterior odds ratio, b_o is the prior odds ratio, and the p_i's and q_i's are the elementary likelihoods for $s_i = 1$ given that the disease actually is $d = 1$ and $d = 0$, respectively. Taking logs of both sides we get

$$\begin{aligned}\ln Y(s) &= \sum_{i=1}^{n} \{(\ln p_i - \ln q_i)\, s_i + (\ln(1-p_i) \\ &\quad - \ln(1-q_i))(1-s_i)\} + \ln b_o \\ &= \sum_{i=1}^{n} \{(\ln p_i - \ln q_i - \ln(1-p_i) + \ln(1-q_i))\, s_i \\ &\quad + (\ln(1-p_i) - \ln(1-q_i)\} + \ln b_o.\end{aligned}$$

Finally, defining $y(s) = \ln Y(s)$ we get the very tidy linear relationship

$$y(s) = w_0 + \sum_{i=1}^{n} w_i s_i, \tag{A1}$$

where

$$w_0 = \ln b_o + \sum_{i=1}^{n} (\ln(1-p_i) - \ln(1-q_i)) \tag{A2}$$

and

$$w_i = \ln p_i - \ln q_i - \ln(1-p_i) + \ln(1-q_i). \tag{A3}$$

for $i = 1, ..., n$. The objective or true values of the internal parameters w_i's specified in (A2) and (A3) are initially known by the experimenter, not by the subjects. Thus the normative learning task can be expressed as finding the $(n+1)$ unknown numerical values $w_0, w_1, ..., w_n$ in the linear relationship (A1) for the log posterior odds $y(s)$.

A comparison to the CMP5 presented in the text reveals that (A1) is precisely the activation arriving at the output node in the CMP5. Recall that the output activation in CMP is transformed using the sigmoid (or logistic) function $S(y) = 1/(1+e^{-y})$ and note that this is precisely the function that transforms log odds to probabilities. Hence the asymptotic CMP can be interpreted as an $n+1$ free parameter version of the Bayesian model, or, equivalently, the Bayes model is a constrained CMP with the parameter restrictions given in Eqs. (A2) and (A3).

The FLMP has a similar interpretation. The $2n$ parameter version presented in the text has no separate base rate parameter b_o, but it is clear that applying the same manipulations (taking logs and collecting terms) that produced (A1) from Bayes will also produce (A1) from $2n$ parameter FLMP, but without the restrictions (A2) and (A3). Thus in effect the parameter b_o can be absorbed into the parameters f and g (or the ratios u and v defined in the text) without affecting the asymptotic model fit. (It turns out that the trial by trial fit of the FLMP version with a separate b_o is inferior because subjects appear to use a slower learning rate λ for this internal parameter than for other internal parameters.)

The literature contains several other connectionist models besides our basic $n+1$ input node model CMP5. Gluck and Bower (1988) investigate the asymptotic version of an $n = 4$ input node (plus one output node) model for the standard n-symptom medical diagnosis task. Their Eq. (6), the central equation for the model, can be written in our notation as

$$y = \sum_{i=1}^{n} w_i s_i,$$

where the connection weights w_i are regression coefficients for the binary independent variables s_i and the dependent variable y is the net activation at the output node. (See their Fig. 2; we used the restrictions given on their page 231 that $\lambda = \alpha = 1$.) This expression clearly would be equivalent to (A1) if an $(n+1)$th input node, constantly at unit activation, were included to provide the missing w_0 parameter as in our CMP5. As for the interpretation of y, Gluck and Bower's Eq. (7) specifies that $P[d = 1 \mid s] = (1+e^{-\theta y})^{-1} = s(\theta y)$, so θy is the log posterior odds. If the parameter θ in a regression can be regarded as a rescaling of all w_i's, then it is inessential in this application. Alternatively, θ can be an additional free parameter used *after* estimating the w_i's, to fit subjects' tendency to respond as if understating or overstating the true odds. Gluck and Bower (1988, p. 234) adopt the free parameter interpretation.

Estes *et al.* (1989) tested two other versions of the connectionist model for a four-symptom ($n = 4$) medical diagnosis task. The first version, call it CMPn + 0, differs from our CMP5 in two respects: (1) two output nodes are used, one for $d = 1$ and the other for $d = 0$; (2) an additional input node 0 is activated if and only if all of the other four nodes are not activated (i.e., $s_i = 0$, $i = 1, ..., 4$). Node 0 is required because otherwise no activation would be available when all symptoms are 0. The second version, call it 2CMP2n, drops node 0 but adds nodes $n+1, n+2, ..., n+n$; node $n+i$ is activated if and only if node i is not active.

It turns out that these connectionist versions differ as learning models but are very similar as asymptotic models. Indeed, 2CMP2n and CMP5 are asymptotically equivalent. To demonstrate, let v_{ij} denote the connection weights for the 2CMP2n model, so it can be written

$$y_j = \sum_{i=1}^{2n} v_{ij} s_i, \qquad j = 1, 0, \tag{A4}$$

where $s_{n+1} = 1 - s_i$, for $i = 1, ..., n$.

Note that the odds ratio for 2CMP2n is $\exp(\theta y_1)/\exp(\theta y_0) = \exp(\theta(y_1 - y_0))$, so we identify y in Eq. (A1) with $y_1 - y_0$ in (A4). Then (A4) becomes

$$
\begin{aligned}
y &= \sum_{i=1}^{2n} (v_{i1} - v_{i0})\, s_i \\
&= \sum_{i=1}^{n} (v_{i1} - v_{i0})\, s_i + (v_{i+n,1} - v_{i+n,0})(1 - s_i) \\
&= \sum_{i=1}^{n} [(v_{i1} - v_{i0} - v_{i+n,1} + v_{i+n,0})]\, s_i \\
&\quad + \left\{ \sum_{i=1}^{n} (v_{i+n,1} - v_{i+n,0}) \right\}.
\end{aligned}
$$

For $i = 1, ..., n$, set w_i equal to the expressions in square brackets [ ] and set w_0 equal to the expression in curly braces { }, and we recover Eq. (A1). Conversely, it is clear from the last expression that there are many ways to go from the CMP5 parameters w_i to the 2CMP2n parameters v_i.

2CMPn + 0 and 2CMP2n are almost asymptotically equivalent. The linearizations of the models coincide, but the models differ in the case when all s_i are 0. To see this, first note that we can again take y as the difference $y_1 - y_0$ to express 2CMPn + 0 as

$$y = \sum_{i=1}^{n} v_i s_i + v_0 \prod_{i=1}^{n} (1 - s_i). \tag{A5}$$

The coefficient v_0 is relevant only when the product in (A5) is nonzero, which is precisely when $s_i = 0$ for all symptoms i. In this case, the last equation reduces to $y = v_0$, that is to say, (A1) with $w_0 = v_0$. When any s_i is nonzero, the equation reduces to $y = \sum_{i=1}^{n} v_i s_i$, that is, (A1) with $w_0 = 0$.

Estes *et al.* (1989) present an exemplar model that assumes that each symptom configuration is stored in memory with fixed probability β, with the category label $d = 1$ or $d = 0$ attached. In the remaining $1 - \beta$ proportion of the trials, there is no memory storage. The similarity of a new symptom configuration s to a stored configuration s^* that matches Y ($s_i^* = s_i$) on exactly $n - k$ symptoms and mismatches ($s_i^* = 1 - s_i$) on exactly $k = K(s, s^*) = \sum_{i=1}^{n} |s_i - s_i^*|$ symptoms is assumed to be α^k, where α is a fixed numerical value between 0 and 1. The model computes the similarity of the new stimulus s each symptom configuration in D_1, the set of stored $d = 1$ stimuli, and sums these computed similarity numbers. Thus the overall similarity index of s for $d = 1$ is given by the formula

$$R_1(\mathbf{s}) = \sum_{\mathbf{s}^* \in D_1} \alpha^{K(s, s^*)} \tag{A6}$$

The overall similarity index $R_0(s)$ of the stimulus s to the alternative ($d = 0$) disease is computed by the same formula except that the sum is taken over s^* in D_0, the set of stored $d = 0$ stimuli. The model uses these similarity indexes and the relative goodness rule (RGR) to predict subjects' responses. That is, the probability of a $d = 1$ response is $R_1/(R_1 + R_0)$.

As the similarity parameter $\alpha \to 0$ the Estes *et al.* (1989) model reduces to our basic exemplar model because the final probability is simply the fraction of stored s-exemplars that are labelled $d = 1$, and that fraction is w_j under the coding used in our basic exemplar model. Fits of the Estes *et al.* (1989) model to medical diagnosis data yield α estimates of 0, so our simple exemplar model apparently loses nothing crucial. (However, we use the standard learning algorithm which produces better trial-by-trial fits than the learning algorithm of Estes *et al.*, 1989).

Nosofsky (1990) presents a related model called the "generalized multiplicative similarity prototype" (GMSP) model. GMSP is a restriction of Eq. (A6) in that the storage probability β is assumed to be 1 and the sets D_0 and D_1 are replaced by their centroids (feature means); this "prototype" restriction gives more influence to outliers and omits the implicit sensitivity to base rates in (A6) when $\#D_1 \neq \#D_0$. On the other hand, GMSP generalizes Eq. (A6) by allowing α to assume different values α_i for different symptoms.

Nosofsky (1990) derives an $n + 1$ parameter version of the GMSP model as follows. Let r_{ij} denote the ratio of similarity constants (α_i for $d = 1$ in the numerator and α_i for $d = 0$ in the denominator) for symptom $i = 1, ..., n$ when the symptom has one value ($j = 1$) or the other ($j = 0$). Then define r_i' as $\sqrt{r_{i1}/r_{i0}}$, and define W_0 to be the product of a bias (i.e., base rate or prior odds) factor and some other terms involving square roots of the r parameters. Then (using Nosofsky's Eqs. A6–A8 and present notation), we can write the posterior odds ratio as

$$\frac{P[d = 1 \mid s]}{P[d = 0 \mid s]} = W_0 \prod_{i=1}^{n} (r_i')^{2s_i}. \tag{A7}$$

If we take logs of both sides of (A7) we once again recover (A1), with $w_0 = \ln W_0$ and $w_i = 2 \ln r_i'$. Hence this version of the GMSP is also asymptotically equivalent to CMP.

To summarize, there are many behavioral models in the psychological literature but they yield only a few observationally distinguishable asymptotic models for the medical diagnosis task. The linear reduced form in Eq. (A1) subsumes the CMP5, 2CMP2n, GMSP, and FLMP. These models are normatively correct in the sense of information integration; we recover Bayes model if we impose the restrictions (A2–A3) that follow from optimal feature evaluation. The other models are related but not equivalent to (A1). The models all assume the relative goodness rule RGR at the decision stage, which may or may not be optimal, depending on the subject's goals.

Learning Algorithms

Learning algorithms specify how the feedback changes the internal parameter values. The learning algorithms we consider can be expressed in the form

$$v_t = v_{t-1} + \lambda_t e(d_t, v_{t-1}), \tag{A8}$$

where v denotes the internal parameter being learned, λ denotes the learning rate, and e denotes the perceived error, given the current feedback d_t and the current parameter value. Our behavioral models all used the same learning algorithm, which goes back to Rescorla and Wagner (1972): a constant learning rate $\lambda_t = \lambda$ and the error specification $e = (d - p(d|s))\, v$—compare Eqs. (5) and (6) of the text, bearing in mind that $S(y)$ is a probability when y is log odds.

The Bayesian learning algorithm is different. Recall that the elementary likelihood of symptom i, given that the target disease is $p_i = [s_i = 1 \,|\, d = 1] = P[s_i = 1 \;\&\; d = 1]/P[d = 1]$. Given observations $t = 1, ..., T$ of s_i and d, the maximum likelihood estimators for the p_i's and the q_i's in the base odds can be expressed in terms of the current counters $N_{\cdot T}$. The estimators are $(1/T)\sum_{t=1}^{T} s_{it}\, d_t = (1/T)\, N_{i\,dT}$ for the numerator of p_i and $(1/T)\sum d_t = (1/T)\, N_{dT}$ for the denominator. Hence $\hat{p}_{iT} = N_{idT}/N_{dT}$ is a consistent estimator for p_i. Similarly, estimate the likelihoods $q_i = P[s_i = 1 \,|\, d = 0]$ by $\hat{q}_{iT} = N_{i\bar{d}T}/N_{\bar{d}T} = (N_{iT} - N_{idT})/(T - N_{dT})$, where $N_{iT} = \sum_{t=1}^{T} s_{it}$, and estimate $b = P[d = 1]/Pr[d = 0]$ by $\hat{b}_T = N_{dT}/(T - N_{dT})$. That is, the appropriate ratios of the current counters $N_{\cdot T}$ are used to estimate the likelihoods and base odds.

To express adjustment to the estimates in terms of learning rates, we solve for the recursions as

$$\begin{aligned}\hat{p}_{iT+1} &= \frac{N_{idT+1}}{N_{dT+1}} = \frac{N_{idT} + s_{iT+1}\, d_{T+1}}{N_{dT+1}} \\ &= \frac{N_{idT}}{N_{dT}} \frac{N_{dT}}{N_{dT+1}} + \frac{s_{idT+1}\, d_{T+1}}{N_{dT+1}} \\ &= \hat{p}_{iT}(1 - d_{T+1}\lambda_{T+1}) + \lambda_{T+1}\, d_{T+1}\, s_{iT+1},\end{aligned}$$

using the fact that $N_{dT} = N_{dT+1} - d_{T+1}$ to get the first term in the last equation, where $\lambda_{T+1} = 1/N_{dT+1}$. Rearranging the last expression, we get

$$\hat{p}_{iT+1} = \hat{p}_{iT} + \lambda_{T+1}\, d_{T+1}(s_{iT+1} - \hat{p}_{iT}), \tag{A9}$$

where again $\lambda_T = 1/\sum_{t=1}^{T} d_t = N_{dT}^{-1}$. Similar calculations yield

$$\hat{q}_{iT+1} = \hat{q}_{iT} + \bar{\lambda}_{T+1}(1 - d_{T+1})(s_{iT+1} - \hat{q}_{iT}) \tag{A10}$$

and

$$\hat{b}_{T+1} = \hat{b}_T + \bar{\lambda}_{T+1}(d_{T+1} - \hat{b}_T + d_{T+1}\hat{b}_T) \tag{A11}$$

for $\bar{\lambda}_T = 1/\sum_{t=1}^{T}(1 - d_t) = (T - N_{dT})^{-1}$.

We conclude that normative Bayesian learning uses a different error function from Rescorla and Wagner and a different learning rate. The learning rate depends on history to some extent; on average $N_{dt} = T \cdot P[d = 1]$, but there is some "sampling" variance about the mean. The most important difference is that the learning rate is not constant, but decreases approximately as $1/T$. In retrospect, this is intuitively obvious: if each observation is equally informative, then the learning rate λ optimally should give equal weight to each current and past observation. Because the expected number of past observations is proportional to T, the learning rate should decline inversely with T.

These conclusions concern learning the $2n + 1$ values p_i, q_i, b of the normative model. Linear recursions do not seem possible for the reduced set of $(n + 1)$ parameters $w_0, ..., w_n$ in Eq. (A1), and logit or probit regressions may be required to obtain efficient estimators. Nevertheless the $\hat{w}_i$'s can be defined in terms of the $\hat{p}$ and $\hat{q}$ expressions and the deceleration rate conclusion still appears valid. That is, $\Delta\hat{w}_i$ is $O(T^{-1})$, not $O(T^{-2})$ or $O(1)$ as seems possible at first given $\exp(\hat{w}_i) = (N_{id})(1 + N_{id} - N_i)/(1 - N_{id})(N_i - N_{id})$.

It is perhaps worth noting that there are tasks for which the normative model calls for a constant learning rate. For example, suppose that the objective symptom likelihoods p_i and q_i are known to drift over time. Under appropriate technical assumptions, Kalman filter techniques (see Meinhold & Singpurwalla, 1983, for example) provide optimal estimates of the parameters w_i of the form

$$\hat{w}_{it} = \hat{w}_{it-1} + \lambda_o(T_i(d_t) - \hat{w}_{it-1}), \tag{A12}$$

as in our basic learning algorithm. The intuition is that when the true numerical values drift, old evidence becomes obsolete at a rate that matches its accumulation rate, so the marginal value of new evidence remains constant at some value λ_o.

REFERENCES

Castellan, N. J., Jr. (1974). The effect of different types of feedback in multiple-cue probability learning. *Organizational Behavior and Human Performances*, **11**, 44–64.

Castellan, N. J., Jr. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, 117–147). Hillsdale, NJ: Lawrence Erlbaum.

Cramer, J. S. (1986). *Econometric applications of maximum likelihood methods*. New York: Cambridge Univ. Press.

Estes, W. K. (1959). The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of science* (Vol. 2, pp. 380–491). New York: McGraw-Hill.

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, & Cognition*, **15**, 556–571.

Gluxk, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 225–244.

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

Massaro, D. W. (1969). A three state markov model for discrimination learning. *Journal of Mathematical Psychology*, **6**, 62–80.

Massaro, D. W. (1987). *Speech perception by ear eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Earlbaum.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97** (2), 225–252.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **8**, 37–50.

Meinhold, R. J., and Singpurwalla, N. D. (1983). Understanding the Kalman filter. *The American Statistician*, **37**, (2), 123–127.

Myers, J. L., & Cruse, D. (1968). Two-choice discrimination learning as a function of stimulus and event probabilities. *Journal of Experimental Psychology*, **77**, 453–459.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, **34**, 393–418.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psycology: Learning, Memory, and Cognition*, **18**, 211–223.

Oden, G. C. (1978). Semantic constraints and judged preference for interpretations of ambiguous sentences. *Memory & Cognition*, **6**, 26–37.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Balck & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton–Century Crofts.

Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science*, **3**, 143–162.

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 613–625.

Wallsten, T. S. (1971). Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology*, **88**, 31–40.

Yellott, J. I., Jr. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, **15**, 109–144.

Received: March 29, 1993