aepapv2

12/29/99

Overconfidence in Judgment for Repeatable Events

Winston R. Sieck **J. Frank Yates**

University of Michigan

Address Correspondence To:
Winston Sieck
University of Michigan
Department of Psychology or Statistics and Management Science
4441 East Hall D3274 Business Admn
Ann Arbor, MI 48109-1109 Ann Arbor, MI 48109-1234

Phone: (734) 615-4584
E-mail: sieck@umich.edu

ABSTRACT

People are overconfident in their judgments about repeatable events. In the current study, a neural network-based probability judgment (NBPJ) model and an exemplar-based probability judgment (EBPJ) model were developed for these kinds of tasks, and accounts for overconfidence were derived from each. The NBPJ asserts that people's learning of ecological probabilities is essentially veridical. However, their classification responses are fundamentally probabilistic, which results in overconfidence. The EBPJ proposes that people learn by storing past examples, and that their judgments are often based on the first example they happen to retrieve. In this model, reliance on small samples of exemplars in judgment leads to overconfidence. Three experiments uniformly supported the EBPJ. Implications for current theories of likelihood judgment are discussed.

Overconfidence in Judgment for Repeatable Events

Consider the following medical diagnosis problem:

Patient: KM

The patient presents with: Rash, Earache

The patient does NOT exhibit: Swollen Hands

Diagnosis (Indicate one): Trebitis Philiosis

Probability that this patient has the indicated disease (50%-100%): _____ %.

As implied by the form of the question, the respondent makes two judgments. First, the respondent states an opinion as to which of the two (hypothetical) diseases is vexing patient KM. The respondent then specifies a probability judgment that the patient actually has the selected disease. Furthermore, once these judgments have been rendered, the respondent learns the truth about patient KM's condition, hence, providing the opportunity to improve his or her future diagnoses.

Previous research with tasks like this one has revealed an interesting overconfidence phenomenon. Specifically, subjects have been found to report average probability judgments that far exceed the proportions of accurate diagnoses that are made (Yates, Lee, Shinotsuka, Patalano, & Sieck, 1998). For instance, in one experiment, American subjects who had previous experience with 60 patients diagnosed another 60 patients. The mear of the probability judgments for this second set of patients was 80%, whereas the percentage of correct diagnoses was only 68%.

Overconfidence in this sense has been an extremely reliable finding in probability judgment tasks using general knowledge or "almanac" questions following this form: "Which species has a longer gestation period: (a) chimpanzees, or (b) humans?" In typical experiments, a respondent first states which of the two alternatives is correct, and then specifies a probability between 50% and 100% that he or she is, in fact, correct (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). Overconfidence in general knowledge is currently a topic of immense interest, and numerous proposed accounts are being vigorously debated (see, for example, the recent reviews by Keren, 1997; McClelland & Bolger, 1994; and Yates, Lee, Shinotsuka, & Sieck, 1999). But it is questionable as to whether the sources of overconfidence in general knowledge tasks are the same as those of the overconfidence found in situations like the medical diagnosis task described above. This is especially true since the situations differ in seemingly important ways. For example, judgments in the medical diagnosis task are concerned with repeatable events; in the sense that they are about the same diseases for every patient. Furthermore, there is a potential for learning in these scenarios when the judgments are also followed by immediate feedback. Since choice behavior under these conditions has been intensively investigated in studies of classification learning, ideas from that domain were expected to provide a useful theoretical point of departure for the current work.

Why is overconfidence in judgments about repeatable events so important? First of all, it is important to the development of judgment theories. As mentioned above, empirical investigations of overconfidence have most often relied on almanac items, and so findings from those kinds of tasks have been especially, perhaps overly, influential on theories of judgment. Overconfidence in repeatable events tasks also poses a challenge to fundamental theory about judgment. Any theory that endeavors to be a universal account for overconfidence must explain its presence in repeatable events tasks as well as in those concerning general knowledge. An unsavory alternative is that we will have to live with many, task-specific accounts of overconfidence.

There is also the practical consideration of the impact overconfidence can have on decision quality, and the very real costs that can be incurred. For instance, in the early 1970's, Royal Dutch/Shell noticed that newly hired geologists were overconfident in t!ıeir predictions about the presence of oil or gas, despite having excellent credentials (Russo & Schoemaker, 1992). Specifically, only one or two wells in ten produced when the geologists estimated a 40% chance of finding oil, thus costing the company enormous amounts of money and time. An overconfident person may also fail to realize that some aspects of his or her judgment procedures are deficient, and hence will refuse to use valuable decision aids. A study by Arkes, Dawes, and Christensen (1986) provides a case in point. These researchers first assessed subjects' knowledge of baseball rules and regulations, and then asked the subjects to indicate which of three baseball players had won the MVP award for each of 20 years. The researchers provided the subjects with the following information for each player from the respective year: batting average, number of home runs, number of runs batted in, and the position of the player's team in the standings. They additionally provided subjects with a quite useful decision rule; the subjects were told that if they always chose the player whose team finished highest in the standings, they would get about 75% of the trials correct. The researchers found that the subjects who knew less about baseball relied more on the rule, and so actually outperformed those subjects who knew more. The more knowledgeable group was, however, more confident in their performance. These results suggest that the more knowledgeable group's overconfidence impaired the quality of their decisions.

Research Strategy

The goal of the present research was an improved understanding of overconfidence in judgments about repeatable events. There is good reason to believe that many factors are capable of affecting overconfidence. The current focus, however, was on contributors that arise from various aspects of the classification learning process. The theoretical approach was as follows. First, two relatively simple, and quite dissimilar, models of classification learning were extended to account for probability judgments. Then, the extended models were scrutinized to ascertain how overconfidence would arise from each. These distinct proposals for repeatable events overconfidence[1] were then subjected to empirical investigation. Specifically, experiments were devised wherein the ordinal predictions of the models differed so as to obtain qualitative tests between them. Since a model that incorrectly predicts the order of effects necessarily achieves a worse goodness of fit than a model that correctly predicts order, regardless of the particular statistic or method used, this qualitative approach yields a much sharper test between the models than would a quantitative approach.

The models adopted for extension were a simple connectionist network model (Gluck & Bower, 1988; Shanks, 1990) and an exemplar retrieval model (Medin & Schaffer, 1978; Nosofsky, 1986). The simple network model essentially assumes that people learn relationships between cues and categories much as would a normal linear regression model (cf. Stone, 1986). In contrast, the exemplar retrieval model suggests that people store individual examples in memory, and base their judgments primarily on the first exemplar retrieved. Of course, there are many other reasonable candidate models of category learning that could potentially be adapted to the current task too (e.g., Anderson, 1991; Ashby, 1992; Gluck, 1992;

Hintzman, 1986; Kruschke, 1992).[2] The connectionist network and exemplar retrieval models were chosen for initial comparison because their conceptions of the categorization process are radically different, elaborating each to account for confidence in classifications is fairly straightforward, and their relative simplicity facilitates the determination of which key assumptions yield overconfidence.

The plan of the rest of this article is as follows. First, each of the models under investigation is described in detail, including the specific accounts for overconfidence implied by each. A series of experiments that test each model's explanation for overconfidence is then reported. In the concluding section, some remaining issues are addressed, and theoretical implications are discussed.

Network-Based Probability Judgment Model

Connectionist network models of category learning were developed and popularized by researchers such as McClelland and Rumelhart (1985), Gluck and Bower (1988), and Shanks (1990). The essential idea of the simplest versions is that a judge learns associations between available cues and categories in a manner analogous to specific proposals for classical conditioning (Rescorla & Wagner, 1972). Suppose, for concreteness, that the judge is a physician faced with the task of diagnosing each of a series of patients as having one of two diseases, Trebitis or Philiosis, as in the example presented earlier. Then the categories are the respective diseases, and the cues might consist of several symptoms, each of which is either present or absent. As the doctor gains experience by observing the symptom patterns of many patients and eventually finding out whether each had Trebitis or Philiosis, she comes to form degrees of association between each of the symptoms and the diseases. The strength of association for each symptom reflects its predictive validity, and can be thought of as that symptom's weight in a regression equation. And the direction of the association, i.e., towards Trebitis or Philiosis, can be represented as a positive or negative sign, respectively, on the regression weight. When the doctor sees a new patient, the individual association strengths for each of the symptoms are combined to form a total degree of association for one of the diseases over the other, say Trebitis over Philiosis. These models typically assume that the doctor's diagnosis depends of this total degree of association in a probabilistic fashion.

That is, the probability that the doctor chooses Trebitis increases with the total degree of association, rather than the doctor always choosing Trebitis once some threshold is exceeded.

The present network-based probability judgment (NBPJ) model is based on ideas described by Gluck and Bower (1988), extending them to account for confidence judgments, reported in the form of probabilities. Suppose a respondent is required to classify a hypothetical patient as having either Trebitis or Philiosis, and then report a 50%-to-100% confidence judgment. According to the NBPJ, this is what happens (see Figure 1; and a formal instantiation of the model is described in Appendix A):

Step 1-Symptom Activation: Mental representations of the symptoms with which the patient presents become activated or "turned on."

Step 2-Category Activation: The association strengths for each of the activated symptoms are then totaled. This total activation is the respondent's internal likelihood that the patient has Trebitis, as opposed to Philiosis.[3]

Step 3-Probabilistic Choice: The chance that the respondent chooses Trebitis, rather than Philiosis, is equivalent to the internal likelihood of Trebitis generated in Step 2.

Step 4-Probability Judgment: The probability judgment that is reported by the respondent depends on whether or not the more likely disease was chosen in Step 3, as follows:

Step 4a-Optimal Choice Case: If the respondent chose the disease indicated to be more likely in Step 2, then the reported probability judgment is a direct reflection of the internal likelihood that the patient has that disease.

Step 4b-Non-Optimal Choice Case: If the respondent chose the disease indicated to be less likely in Step 2, then the reported probability judgment is adjusted to accommodate the 50%-100% scale, and increases with the respondent's internal likelihood that the choice made in Step 3 is correct.

Step 5-Learning: The respondent's association strengths between the symptoms and diseases are adjusted to more appropriate values in light of what disease the patient is actually found to have.

/ Insert Figure 1 about here /

The assumption of probabilistic choice in Step 3 is key to the NBPJ's account of overconfidence.[4] The intuition is that the respondent has less than a 50% chance of being correct for those trials on which he or she predicts that the less likely event will occur. However the respondent's reported probability that the diagnosis is correct is constrained to be at least 50%. Hence, these non-optimal choice trials contribute considerable overconfidence to the respondent's judgments. The probabilistic choice assumption was made, but not extensively discussed in earlier network models. It is addressed here because of its key role in the NBPJ's account of overconfidence.

The empirical basis for assuming that choice is probabilistic is the classical phenomenon of probability matching (Grant, Hake, & Hornseth, 1951; Humphreys, 1939). That is, for example, when an event A actually occurs 70% of the time, people tend to predict the occurrence of event A on 70% of the trials, rather than on every trial. This effect, at least to a first approximation, has been replicated in innumerable studies (cf. Estes, 1964).[5] The learning rule typically used by adaptive network models would, in the long run, lead to an activation level for the event A that is equivalent to the actual proportion of times that A occurs. This is because the rule is designed to minimize the expected mean squared error between activation level and target event occurrence (Gluck & Bower, 1990). So, if learning happens after that fashion, then there must be a rule that takes the degree of activation as input, and probabilistically returns a response.

Another reason to propose that a probabilistic response rule exists is that there are, in fact, situations where such a rule would be decidedly advantageous. For example, it is useful when one is trying to predict the actions of an intelligent adversary. In this case, the adversary will undoubtedly change his strategy if you predict his most likely action in every encounter (and interestingly, in some early studies of probability learning, subjects were instructed to "outguess the experimenter on each trial," e.g., Estes & Straughan, 1954). Analyses from game theory imply that the optimal solution in these situations is to respond to the adversary's most likely action with the same probability that that action occurs (e.g. Simon, 1956). Hence, it could be that such a rule is induced from many previous experiences, and then overgeneralized to all kinds of prediction tasks, such as the one under consideration here.

Exemplar-Based Probability Judgment Model

Exemplar models of category learning were originated by Medin and Schaffer (1978), and have been studied extensively since then (cf. Medin, 1992; Nosofsky, 1992). The essential idea is that a judge learns by accumulating distinct experiences of cues and categories; in the physician example, storing the symptom patterns and disease outcomes of patient after patient. When the doctor receives a new patient, similar, previously encountered, exemplars are brought to mind and used to arrive at a diagnosis. Quantitative descriptions of exemplar models often represent the process as one in which similarity between the new patient and all existing exemplars is computed, and these individual similarity values are combined to form a total degree of similarity between the new patient and one of the diseases over the other, say Trebitis over Philiosis. The doctor's diagnosis is represented as a probabilistic response that depends on the total degree of similarity. However, Medin and Shaffer (1978) described an alternative conception of the process, whereby the doctor's diagnosis, in our ongoing example, is often based on the first exemplar retrieved. Under this interpretation, computations are not literally performed on all existing exemplars; instead, the formulae present a shorthand for indicating the probability that a retrieved exemplar will be associated with one of the disease categories. And this essential idea is also consistent with other current exemplar models (e.g. Nosofsky & Palmeri, 1997; Smith, Patalano, & Jonides, 1998). Hence, this latter interpretation forms the basis for development of the current model.

The present exemplar-based probability judgment (EBPJ) model elaborates on the ideas of Medin and Shaffer (1978) to account for probability judgments in addition to classification behavior. Suppose a respondent is required to classify a hypothetical patient as having either Trebitis or Philiosis, and then report a 50%-to-100% probability judgment that the classification was correct. This is what happens according to the EBPJ (see Figure 2; a formal instantiation of the model is described in Appendix B):

Step 1-Exemplar Retrieval: One after another, the respondent retrieves recent, similar past instances, each of which indicates either Trebitis or Philiosis.

Step 2-Balance Assessment: The respondent assesses the extent to which, on balance, the collection of exemplars retrieved in Step 1 supports either Trebitis or Philiosis.

Step 3-Choice: The respondent chooses the disease favored in Step 2.

Step 4-Probability Judgment: The respondent reports a probability judgment that the patient actually has the disease indicated in Step 3 according to the magnitude of the balance assessment from Step 2.

Step 5-Learning: The respondent may store in memory the symptoms and disease the patient is actually found to have.

Key Assumption 1-Attenuated Responsiveness: The first exemplar retrieved has the greatest impact on the balance assessment. As retrieval proceeds, responsiveness to new exemplars decreases.

Key Assumption 2-Abbreviated Retrieval: Exemplar retrieval tends to terminate after an exceedingly small number of cycles.

/ Insert Figure 2 about here /

The key assumptions jointly imply that judgment is often based on the first exemplar retrieved, and they are critical to this model's account of overconfidence.[6] Suppose that the first retrieved exemplar has an extreme impact on balance. The respondent will then be exceedingly confident in the conclusion that exemplar affords. The only way that confidence can be reduced, or the alternative conclusion reached, is for the process to continue so that exemplars favoring the alternative conclusion have some chance of being brought to mind. However, Assumption 2 implies that little, if any, additional retrieval will occur. There will also be more sampling variability in the balance assessments when fewer exemplars are retrieved, and greater variability in such internal likelihoods has been shown to imply larger degrees of overconfidence (Juslin, Olsson, & Björkman, 1997; Soll, 1996). Because these assumptions are so important to the EBPJ's account of overconfidence, they are further elaborated below.

There are several reasons we should expect responsiveness to be attenuated and retrieval to be abbreviated. The first reason to expect attenuated retrieval is based largely on work in impression formation. In the standard procedure, a few personality-trait adjectives for some hypothetical person are presented sequentially, and then subjects judge the likableness of that person. A very reliable effect in these studies is that the earlier adjectives are most influential, a reasonable account being that less attention is given to the later adjectives (e.g. Anderson, 1973). Attention decrement should also operate at the balance assessment stage, since people have the option to stop attending (Hogarth & Einhorn, 1992).

A second argument is related to the availability heuristic (Tversky & Kahneman, 1973). The idea is that people tend to assign special status to "spontaneously" retrieved memories, which are more likely to occur very early in the retrieval process. According to the intuition, the person reasons: "That memory must be especially relevant; why else would *it* have come to mind?"

A reason to expect abbreviated retrieval is that it is effortful, at least to some degree. Research on contingent use of strategies in decision making (cf. Payne, Bettman, & Johnson, 1992) implies that people will be inclined to terminate retrieval quickly, so as to minimize the cost associated with retrieval effort. Also, people have been shown to believe in a "Law of Small Numbers" (Tversky & Kahneman, 1971). That is, they tend to draw stronger conclusions from limited amounts of data than are warranted by the normative principle of the Law of Large Numbers. People do not retrieve larger amounts of evidence because they see no need to do so.

The final reason for anticipating abbreviated retrieval is that it can be adaptive in certain types of ecologies, when based on recent events. For example, disease is often epidemic, and people nearest to the source are generally the first to contract it. Such a process will lead to high degrees of spatial autocorrelation in the afflictions people suffer. For the local doctor who is visited by one patient after another, that spatial autocorrelation is experienced as temporal autocorrelation. In situations like this, the doctor's diagnoses might suffer if they are based on balance assessments taken over all past experiences. It pays to sample recent events in autocorrelated environments. Data on the decision behavior of bumble bees adds further support to this third reason. Leslie Real and his colleagues have conducted a number of interesting studies on animal choice behavior, using the bumble bee as a "model system" (Real, 1991). In one study, Dukas and Real (1993) found that bees base their foraging patterns (flight distance to next flower) primarily on the nectar reward from the last flower encountered. Specifically, they were found to fly shorter distances after visiting flowers with nectar than without nectar. The bees "behaved as if they were foraging over a patchily distributed resource" (p. 246). Real (1991) argued that nectar rewards in wild fields are often characterized by high degrees of spatial autocorrelation, and that relying on a few recent events is advantageous under such conditions.

This reason for abbreviated retrieval is compelling, but only if it is combined with a recency assumption. And recency effects were generally found in the early probability learning literature (e.g. Estes, 1957). So, as indicated in Step 1, the present rendition of the model assumes that more recent events are more likely to be retrieved than are earlier events that are otherwise equally similar (also see Nosofsky, Kruschke, & McKinley, 1992). This assumption is checked in the empirical studies, to be described next.

Experiment 1: Assessment Method

The previously described two-stage method of eliciting probability judgments is not the only available method. For example, Ronis and Yates (1987) assessed probabilities in two ways. First, they used the previously discussed standard method of requesting the person to choose the answer he or she felt was correct, and then provide a 50%-to-100% probability that the answer was in fact correct. In the second assessment method, items were initially circled (randomly) by the investigators, and then subjects were asked to provide a probability, ranging from 0% to 100%, that the circled alternative was correct. The former method was dubbed the Choice-50 (C50) procedure, and the latter was called the No-choice-100 (NC100) procedure by Ronis and Yates.

The first experiment here varied the assessment procedures in very much the same way as in the Ronis and Yates research. The NBPJ and EBPJ make opposing predictions about how overconfidence will differ in these two procedures, thus providing a direct test. Specifically, the NBPJ predicts that overconfidence will be reduced for the NC100 task, whereas the EBPJ predicts that it will be amplified. These predictions are described, as follows.

According to the NBPJ, a relatively constant internal probability arises from repeated exposure to a cue ensemble, and then a probabilistic response rule is applied. Because responses are probabilistic in the C50 task, the respondent sometimes selects the less likely option. In these cases, the respondent has less than an even chance of being correct, but must report a confidence level of at least 50%. Thus, such trials cause the respondent's judgments to exhibit overconfidence on the whole. But what if the judge's task is to directly report a probability that the patient has Trebitis, without first indicating a choice, as in the NC100 procedure? This procedure eliminates the choice process, effectively bypassing the probabilistic response. The process for the NC100 task would thus be as follows (see Figure 3):

Step 1-Symptom Activation: Mental representations of the symptoms with which the patient presents become activated or "turned on."

Step 2-Category Activation: The association strengths for each of the activated symptoms are then totaled. This total activation is the respondent's internal likelihood that the patient has Trebitis, as opposed to Philiosis.

Step 3-Probability Judgment: A probability judgment that the patient has Trebitis is reported based on the internal likelihood generated in Step 2.

Step 4-Learning: The respondent's association strengths between the symptoms and diseases are adjusted to more appropriate values in light of the disease the patient is actually found to have.

/ Insert Figure 3 about here /

Since overconfidence results largely from the probabilistic response rule, bypassing the stage in which the rule applies should substantially reduce overconfidence (see Appendix A). A study by Neimark and Shuford (1959) lends empirical support for this proposal. In a standard probability learning task, those researchers had subjects make predictions regarding a deck of cards that they turned up one after another. One group of subjects predicted which of two letters would appear on each trial. Another group predicted similarly, and also estimated the percentage of cards in the deck that would contain one of the two letters. The "prediction" group's choice proportions closely approximated the probability that the more frequent letter would appear. The "estimation" group's estimates also closely approximated the probability that the more frequent letter would appear. Neimark and Shufford's results suggest that people's choices will coincide with base rates probabilistically, but that their estimates will match them directly (also see Shanks, 1991). One potential confound, however, is that the estimation instruction may have influenced people to bring more of the past outcomes to mind.

The EBPJ's conception corresponds with this latter possibility, and so it makes just the opposite prediction concerning the current experiment. According to this model, judgments are based on small samples of exemplars that have been stored in memory. The exact sample sizes are presumed to vary according to exogenous factors, but more exemplars should be retrieved with more prompting. In the C50 task, there are two implicit demands for the subject to retrieve information, one at the choice prompt, and one at the confidence prompt. In the NC100 task, there is only one such demand for retrieval, since subjects are only prompted for a probability judgment. Hence, the probability judgments made by subjects in the C50 task should tend to be based on larger samples of exemplars than those made by subjects in the NC100 task, so that greater overconfidence is found in the latter group (also see Sniezek, Paese, & Switzer, 1990).

Method

Subjects

Study participants were 86 undergraduate students enrolled in an introductory psychology course at the University of Michigan. Experimental participation was part of their course requirement.

Cover Story and Ecology

The cover story was adapted from that of Yates et al. (1998). Study participants were asked to imagine that they were physicians in the following scenario: Two new diseases have appeared in the physician's community, "Trebitis" and "Philiosis." The physician will see a series of patients, each of whom has or does not have each of three symptoms: runny nose, swollen hands, sore throat. These symptoms are suspected, but not guaranteed, to be useful in distinguishing patients with Trebitis from those with Philiosis. For each case, the physician's task is to formulate a probabilistic differential diagnosis between Trebitis and Philiosis.

Several notational conveniences are adopted in the following description of the ecology: (a) the disease Trebitis is arbitrarily labeled the "focal disease" or, more generally, the "target event"; (b) the symptoms are abbreviated as N = runny nose, H = swollen hands, and S = sore throat.

Contingency table. Subjects saw a total of 140 "patients," divided into two blocks of 70. The same trial sequence was presented to all subjects. Table 1 shows the distribution of symptom patterns and Trebitis outcomes for each block. In this article, the relative frequency of Trebitis for each symptom pattern is referred to as a conditional "objective" probability for the artificial ecology. However, we do not presume that objective probabilities exist in real ecologies.

/ Insert Table 1 about here /

Base rates. The base rates for Trebitis and Philiosis were .50/.50. Each of the symptoms had a base rate of .57.

Cue validities. The validities of the symptoms for distinguishing Trebitis from Philiosis are indicated via ΔP, the statistic often used in contingency judgment research. For each symptom, it is defined as:

$$\Delta P = P(\text{Trebitis}|\text{Symptom Present}) - P(\text{Trebitis}|\text{Symptom Absent}) \qquad (2.1)$$

The validities for the symptoms are: $\Delta P_N = .41$, $\Delta P_H = -.41$, $\Delta P_S = .12$.

Predictability. A logistic regression model was created for the patients in the present ecology. The model was:

$$p = \frac{e^Y}{1+e^Y} \qquad (2.2)$$

where $p = P^*(T)$ was the model's probability "judgment" for the target event (in this case, Trebitis).[7] Y was the following linear combination of indicator variables, coded 1 and 0 for the presence and absence of the symptoms, respectively:

$$Y = -0.65 + 1.95\text{N} - 1.63\text{H} + .76\text{S} \qquad (2.3)$$

The proportion of variance explained by the model was $R^2 = .24$. Also, the model's "judgments" achieved 73% correct diagnoses.

Design

The independent variable in the experiment was the method of elicitation of the probabilistic differential diagnosis, and two methods were used (see Figure 4). One was the standard "Choice-50" (C50) method, in which the subject was first asked to indicate the disease he or she felt each patient was most likely to have, and then to provide a 50%-to-100% probability that the patient actually had the indicated disease. The second method was the "No-choice-100" (NC100) assessment procedure, in which subjects were asked to provide a probability, ranging from 0% to 100%, that each patient had the focal disease, Trebitis.

/ Insert Figure 4 about here /

Procedure

The experiment was conducted entirely via computer. The program first introduced the scenario and initial instructions to the subject-"physician." These instructions were very similar to those used by Yates et al. (1998), and emphasized several points, including: (a) the need to learn the relations between the symptoms and diseases over time in order to make good diagnoses; (b) that each symptom may or may not be useful in distinguishing Trebitis from Philiosis; and (c) that the diagnostic process is inherently probabilistic rather than deterministic, unlike what the subject might have expected in a psychology experiment.

Subjects were also given specific instructions concerning use of the probability scale. Subjects in the C50 condition were instructed to adhere to the following conventions when stating their likelihood judgments:

A) 50% should mean that the patient is just as likely to have Trebitis as Philiosis.

B) 100% should mean that the patient is absolutely certain to have the disease you indicated for your first judgment.

C) Increasing probabilities between 50% and 100% should correspond to increasing degrees of certainty that the patient's true medical condition is as you stated.

Subjects in the NC100 condition were told to adhere to these conventions:

A) 50% should mean that the patient is just as likely to have Trebitis as Philiosis.

B) 100% should mean that the patient is absolutely certain to have Trebitis, and 0% should mean that the patient is absolutely certain to have Philiosis.

C) Increasing probabilities between 50% and 100% should correspond to increasing degrees of certainty that the patient has Trebitis rather than Philiosis.

D) Decreasing probabilities between 50% and 0% should correspond to increasing degrees of certainty that the patient has Philiosis rather than Trebitis.

On each trial, the subject: (1) was presented with a new patient, who was identified by two initials, and that patient's symptom profile; (2) indicated a probabilistic differential diagnosis according to either the C50 or NC100 procedure, as described above; and (3) received feedback about what was "eventually determined" to be the patient's actual condition. Subjects made diagnoses for 70 patients during an initial block of trials, and then returned to diagnose another 70 patients in a second block, two hours later. Analyses focus on Block 2, because those judgments are of primary importance for the hypotheses under consideration.

Results and Discussion

In the analyses described below, choices were derived from the NC100 subjects' data by a cutoff rule, such that probabilities greater than 50% were mapped to predictions for the focal disease and those less than 50% were mapped to predictions for the non-focal alternative. Choices were randomly selected for judgments of exactly 50%. Also, probability judgments of the focal disease were derived from the C50 subjects by taking their judgments "as is" when they chose the focal disease, and by subtracting their judgments from 100% when they chose the non-focal alternative.

Proportion Correct and Bias

The proportion correct achieved by the C50 condition subjects was slightly higher than that achieved by the NC100 condition subjects, $M = .602$ and .569, respectively, $t(84) = 2.01, p = .048$. The EBPJ anticipates either this result, or no difference between the conditions, but the NBPJ predicts the opposite result.

Over/underconfidence is typically indexed via the following bias statistic:

$$\textit{Bias} = \textit{mean probability judgment - proportion correct} \qquad (2.4)$$

Positive values for bias indicate overconfidence, and negative values indicate underconfidence. As Figure 2.3 shows, subjects in both conditions exhibited marked positive bias, i.e. overconfidence. Also, the positive bias was much greater in the NC100 condition than in the C50 condition, $t(84) = 3.66, p = .0005$. This result runs counter to the predictions of the NBPJ, but supports the EBPJ's predictions.

/ Insert Figure 5 about here /

Recent Outcomes

A regression analysis was performed to determine the influence of recent outcomes on current judgments. In the analysis, the mean probability judgment of Trebitis was computed for each trial. These mean judgments were then regressed on the corresponding symptom values, indicator variables for the disease outcomes of the previous three trials (denoted D_{-1}, D_{-2}, and D_{-3}, respectively; D_{-i} was 1 if the ith past patient had Trebitis, and 0 otherwise), the number of common symptoms shared by the current trial and each of the previous three trials (denoted C_{-1}, C_{-2}, and C_{-3}, respectively), and the products between the respective disease outcome and symptom commonality terms. It is these product terms that are of particular interest. Specifically, if it is the recent, similar exemplars that are most likely to be retrieved (as argued previously), then positive values for these terms should be expected. The analysis was performed separately for each condition. The percentage of variance explained by the model for the C50 condition was $R^2 = .83$, and the percentage of variance explained for the NC100 condition was $R^2 = .74$. The results are displayed in Table 2, by condition. As can be seen, statistically significant, positive product terms do exist for the most immediate preceding two cases. This result justifies the recency assumption of the EBPJ.

/ Insert Table 2 about here /

Judgment Given Symptom Pattern

Table 3 shows, for each symptom pattern: (a) the objective probability of Trebitis, as derived from the frequencies presented in Table 1; (b) the proportion of times Trebitis was chosen by subjects in each condition; and (c) the mean probability judgments of Trebitis. Choice proportions from both conditions appear to overshoot the objective probability at the lowest values and undershoot at the highest values, rather than match them. The choice proportions are not closer to optimal (i.e. nearer to 0 and 1) for the NC100 condition than for the C50 condition, contrary to the predictions of the NBPJ. The mean probability judgments exhibit the same pattern of overshooting the objective probability at the lowest values and undershooting at the highest values. This pattern is consistent with the notion that judgments are perturbed by random error (cf. Erev, Wallsten, & Budescu, 1994).. However, it is not diagnostic because both models anticipate such error, although by different mechanisms.

/ Insert Table 3 abov: here /

Summary

The findings of Experiment 1 provide some support for the EBPJ's account of repeatable events overconfidence. As proposed by that model, subjects were less overconfident in the C50 task than in the NC100 task. And this finding was exactly opposite to the expectations of the NBPJ, thereby providing evidence against its account of overconfidence. Further evidence against the NBPJ's validity was that choices derived by a cutoff rule from the probability judgments given by subjects in the NC100 condition resembled those of the C50 group quite closely. This finding is contrary to that model's assumption that choice is essentially a probabilistic function of internal likelihood. One other important finding was that recent outcomes played a special role in choice, essentially generalizing positive recency effects found in early studies of probability learning to the multiple-cue case. This finding supports the decision to modify Medin and Schaffer's (1978) retrieval rule, as discussed in the introduction.

Although the Experiment 1 overconfidence results were predicted by the EBPJ, there are difficulties with the interpretation. A limitation is that the C50 condition subjects achieved a higher proportion of correct responses than did the NC100 subjects. This is problematic because of what are often called hard-easy effects (Lichtenstein & Fischhoff, 1977; Suantak, Bolger, & Ferrell, 1996). These effects imply that less overconfidence should be expected when a higher proportion of correct responses is achieved. And the difference in proportion correct found here might well stem from differences in learning. Specifically, the C50 group is prompted to retrieve information more times on each trial than the NC100 group, and learning is often assumed to occur during retrieval (Gillund & Shiffrin, 1984). So, it might be argued that the overconfidence results simply reflect that more learning occurred in the C50 condition.

Another possible alternative to the exemplar retrieval interpretation of the overconfidence result is that there is a linguistic demand to report a more extreme probability in the absence of a categorical judgment. For example, the respondent may feel that stating "Trebitis" along with 60% confidence is more informative than stating a probability judgment of 60% that a patient has Trebitis. If so, the respondent might be compelled by the conversational rule of informativeness (cf. Grice, 1975) to report a probability slightly greater than 60% in the latter case. In principle, this linguistic phenomenon could be operating at all points along the scale. These issues are addressed in Experiment 2.

Experiment 2: Assessment Method and Recall

The results of Experiment 1 support the EBPJ's account of overconfidence, but several interpretational issues remain. Hence, the primary purpose of Experiment 2 was to provide more direct evidence for the EBPJ's explanation for the effect of assessment method on overconfidence. To that end, several modifications were made to the basic design of Experiment 1.

One possible alternative explanation for the assessment effect was differential learning. In Experiment 2, subjects in all conditions simply classified patients without giving confidence judgments in Block 1, and no subjects received feedback during Block 2. Thus, learning conditions were equalized in the present experiment.

Another possibility was that subjects in the NC100 condition felt it necessary to report more extreme probability judgments than did subjects in the C50 condition. This could occur because respondents believe that the latter construction is more informative in conversation than the former. This response bias interpretation was dealt with in two ways. First, a monetary bonus system that encourages respondents to report candidly was included. Second, a "recall" condition was added in order to directly test the exemplar retrieval account. Specifically, after presentation of each patient, subjects in the recall condition were instructed to recall as many similar patients as possible. They were then prompted for a diagnosis, according to the NC100 procedure. If the assessment effect is due to increased retrieval, as suggested by the EBPJ, overconfidence should be reduced for this group, as in the C50 condition. However, this manipulation does not change the conversational demand, since no more information is being communicated than in the NC100 "control" condition.

One other possibility, not previously discussed, is that people simply attend more completely to the symptoms in the C50 condition. As shown in Appendix B, such an attention effect might be expected according to the EBPJ, independent of the specific key assumption that abbreviated retrieval drives overconfidence. And any effect for the recall condition could also be interpreted as attentional in nature. In order to control for this possibility, an "encoding" condition was added wherein subjects were instructed to pay close attention to the symptoms on each trial, in place of receiving the recall instruction or choice demand. Subjects in this condition also made their judgments according to the NC100 procedure. This condition obviously controls for effects of attention.

The EBPJ predicts that the basic assessment method effect will be replicated under the more stringent learning conditions of the present experiment. Furthermore, since subjects in both the C50 and recall conditions were prompted twice to retrieve information, it expects that similar levels of overconfidence will be observed in those groups.

Method

Subjects

Study participants were 159 undergraduate students enrolled in an introductory psychology course at the University of Michigan. Experimental participation was part of their course requirement.

Cover Story and Ecology

The cover story was exactly the same as in Experiment 1. However, the ecology was changed slightly, as described below.

Contingency table. Subjects saw a total of 120 "patients," divided into two blocks. There were 90 patients in Block 1, and 30 patients in Block 2. Table 3.1 shows the distribution of symptom patterns and Trebitis outcomes for Block 1. The Block 2 ecology consisted of a representative sample of the Block 1 ecology.

/ Insert Table 4 about here /

Base rates. The base rates for Trebitis and Philiosis were .50/.50. The base rates for symptoms N and H were .67, and the base rate for symptom S was .50.

Cue validities. The validities of the symptoms for distinguishing Trebitis from Philiosis were: $\Delta P_N = -.50$, $\Delta P_H = .50$, $\Delta P_S = .16$.

Predictability. A logistic regression model describing Block 1 was created according to Equation 2.2. In this case, *Y* was the following linear combination of indicator variables for the presence and absence of the symptoms:

$$Y = -0.48 - 1.62\text{N} + 1.62\text{H} + .96\text{S} \qquad (3.1)$$

The proportion of variance explained by the model was $R^2 = .25$. Also, the percentage of correct diagnoses achieved by the model's "judgments" applied to Block 2 was 73%.

Design

There were four between-subjects conditions in Experiment 2: control, encode, recall, and C50 (see Figure 3.1). All factors were introduced, and no feedback was given, in Block 2. The NC100 assessment method, described in Experiment 1, was employed in the first three conditions. Subjects in the control condition simply made judgments according to the NC100 procedure, as described in Experiment 1. Subjects in the encode condition were presented with the following instruction for 3 seconds at the time the symptom profile was displayed, but prior to making their judgments: "Carefully examine this patient's symptom profile." Subjects in the recall condition were presented with this instruction: "Carefully examine this patient's symptom profile. Try to bring to mind all of the patients you saw previously with symptoms like these, including the disease each had." Subjects in the fourth condition made judgments according to the C50 procedure, as described in Experiment 1.

/ Insert Figure 6 about here /

Procedure

The procedure was very similar to that of the first experiment; the scenario and initial instructions were essentially identical. During Block 1, subjects diagnosed 45 patients, took a short break during which they engaged in an unrelated activity, and then diagnosed another 45 patients, for a total of 90 patients. Subjects in all conditions made only choices during Block 1; no probability judgments were rendered. A second set of instructions was presented just before the start of the Block 2 trials. These instructions informed subjects of any procedure changes they would encounter in their diagnostic routine (i.e. specific condition instructions, and corresponding probability judgment conventions). Subjects in the recall condition were also given the following information about what they might reasonably expect from their recall attempts:

> You may feel that it is difficult or impossible to remember all of the patients you saw exactly, but even though you probably cannot identify them, you *can* remember some bits about them, including which of the two diseases they had.

The subjects were also informed that a formula would be used to evaluate their judgments from the upcoming, final phase, and that their score, along with the average scores of their peers, would be sent to them at the end of the semester. They also learned that the subjects with the four best accuracy scores would each receive a $20 bonus prize. This was intended to encourage effort and accuracy on the subjects' part. The instructions emphasized that the scoring procedure had a special characteristic which implied that it was in the subject's best interest to be perfectly candid in reporting his or her true judgments. The characteristic is technically called "properness," though this term was not used with the subjects (cf. Yates, 1990, chapter 8). In Block 2, subjects diagnosed another 30 patients according to one of the four conditions described above. Both blocks of diagnoses were rendered within a single hour. Analyses focus on the second block, because only those judgments bear on the hypotheses under consideration.

Results and Discussion

Expected Proportion Correct and Expected Bias

Since no subjects received feedback during Block 2 of this experiment, no actual outcome values existed for computing proportion correct and bias. So, the objective conditional probabilities of Trebitis were used to determine the probability that each answer would be correct for a given set of generated outcomes. These values were then averaged to obtain the expected proportion of correct answers. The mean expected proportions of correct answers were essentially identical across conditions, $M = .610$, .601, .610, and .607 for control, encode, recall, and C50 groups, respectively.

Over/underconfidence was indexed via the following expected bias statistic:

$$E[Bias] = mean\ probability\ judgment - E[proportion\ correct] \qquad (3.1)$$

Figure 3.2 shows the mean values of expected bias for each of the four conditions. As shown, overconfidence was nearly equivalent for subjects in the control and encode conditions. Those subjects were more overconfident than subjects in the recall condition, $t(117) = 2.08, p = .040$, and also than subjects in the C50 condition, $t(117) = 2.30, p = .024$. Finally, subjects in the recall and C50 conditions were about equally overconfident. This pattern of results is just as anticipated by the EBPJ.

/ Insert Figure 7 about here /

Judgment Given Symptom Pattern

Table 3.2 shows, for each symptom pattern: (a) the objective probability of Trebitis; (b) the proportion of times Trebitis was chosen by subjects in each condition; and (c) the mean probability judgments of Trebitis by condition. Choice proportions exhibited probability matching to a very rough approximation. As in the first experiment, the mean probability judgments overshot the objective probability at the lowest values, and undershot at the highest values.

/ Insert Table 5 about here /

Summary

In this experiment, the assessment method effect found in Experiment 1 was replicated under more stringent conditions. And the similarity in overconfidence between the recall and C50 conditions increases the plausibility that recall drives that effect. Finally, no reduction in overconfidence was found in the encoding condition, effectively ruling out attention as an explanation.

One alternative explanation of the results that was not explicitly controlled for here suggests that people do not actually follow the retrieval instruction, but the delay that is incurred prompts them to shift to a more analytical mode of deliberation (e.g. Hagafors & Brehmer, 1983). An immediate difficulty for this interpretation is that analytical modes have not generally been associated with reduced overconfidence (e.g. Paese & Sniezek, 1991; Wilson & LaFleur, 1995). For example, Wilson and LaFleur found that analyzing reasons for acting or not acting in a certain way led to both a decrease in predictive accuracy and an increase in confidence that the predicted behavior would occur (also see Sieck, Quinn, & Schooler, 1999). Furthermore, according to the mode proposal, the specific instruction should not be relevant; it should be the associated delay that influences the shift. Since the same delay is used for the encoding condition as for recall, the encoding condition at least serves as an indirect control for the mode explanation.

The pattern of results is quite consistent with the proposal that people store exemplars during learning, and then retrieve only small portions of them in order to arrive at their judgments. Since individual records of past experiences simply do not exist according to the NBPJ, it does not anticipate the recall instruction effect. And as discussed in Experiment 1, that model predicts the opposite effect of assessment method. Hence, Experiment 2 provides no evidence that the NBPJ's explanation of overconfidence is accurate.

Experiment 3: Recall Prior to Choice

In Experiment 2, instructing subjects to recall previously seen patients prior to reporting probability judgments was as effective at reducing overconfidence as an initial prompt for choice. The principal objective of Experiment 3 was to determine whether such an instruction could reduce overconfidence over and above a demand to choose. That is, could combined choice and recall demands reduce overconfidence more than either demand alone? To that end, control subjects in this experiment followed the C50 procedure. Experimental subjects performed the same task, but were also instructed to recall as many similar patients as possible, prior to offering their diagnoses. Including these multiple prompts to retrieve information should provide some idea of how effective mere recall can be as a debiasing strategy, when taken to a limit. Of course, the effectiveness should depend on the extent to which responsiveness to new exemplars is attenuated. If attenuation is minimal, so that retrieved exemplars are responded to equally, then the recall instruction should lead to both an increase in proportion correct, and a reduction in overconfidence. However, if the first exemplar retrieved is the main influence on choice, and subsequent retrieval primarily influences confidence, then the procedure should lead to some reduced overconfidence, but have little or no impact on the proportion of correct answers.

Method

Subjects

Study participants were 56 undergraduate students enrolled in an introductory psychology course at the University of Michigan. Experimental participation was part of their course requirement.

Cover Story and Ecology

The cover story and ecology were exactly the same as in Experiment 2.

Design

The independent variable in this experiment was the recall instruction manipulation, with the factor being introduced in Block 2 (see Figure 4.1). Subjects in the recall condition were presented with the following instruction for 5 seconds at the time the symptom profile was displayed, but prior to making their judgments: "Try to bring to mind all of the patients you saw previously with symptoms like these, including the disease each had." No such instruction was given to subjects in the control condition.

/ Insert Figure 8 about here /

Procedure

The procedure was essentially the same as that of Experiment 2. One difference was that subjects in both conditions made judgments via the C50 procedure during the Block 1 trials, since they would all follow that same procedure in Block 2. A second was that the retrieval instruction in this experiment appeared on a separate screen from the judgments, and was displayed for 5 rather than 3 seconds. These changes were made in an attempt to strengthen the effectiveness of the procedure.

Results and Discussion

Expected Proportion Correct and Expected Bias

The expected proportion correct achieved by subjects who received the recall instruction was not statistically different than that achieved by subjects who did not receive the instruction, M = .637 and .591, respectively.

Figure 4.2 shows the mean values of expected bias for the two conditions. As predicted by the EBPJ, subjects in the recall condition exhibited less overconfidence than those in the control condition, $t(54) = 2.15$, $p = .036$. However, as can be seen in Figure 4.2, overconfidence in the recall condition was still significantly positive, $t(27) = 2.23$, $p = .035$, suggesting that multiple retrieval demands cannot completely eliminate overconfidence.

/ Insert Figure 9 about here /

Judgment Given Symptom Pattern

Table 4.1 shows, for each symptom pattern: (a) the objective probability of Trebitis; (b) the proportion of times Trebitis wa; chosen by subjects in each condition; and (c) the mean probability judgments of Trebitis by condition. Choice proportions exhibited probability matching to a very rough approximation. As in the previous experiments, the mean probability judgments overshot the objective probability at the lowest values, and undershot at the highest values.

/ Insert Table 6 about here /

Summary

Experiment 3 provided more direct support for the proposed mechanisms of exemplar retrieval as contributors to overconfidence. Specifically, an instruction to retrieve many exemplars prior to choice led to a reduction in overconfidence, over and above the effect of assessment method. It also suggested that instructions to retrieve more information than is usual generally will be insufficient for achieving completely unbiased confidence assessments. This is expected by the EBPJ, via the assumption of attenuated responsiveness. Indeed, the pattern of results is, on the whole, well accounted for by the EBPJ. Further issues are discussed in the conclusions chapter.

General Discussion

The results of the three experiments reported here provide encouraging support to mechanisms of exemplar retrieval as contributors to overconfidence in judgments for repeatable events. An NBPJ account of overconfidence was not supported. Experiment 1 showed that overconfidence was greater when judgments were prompted in two stages, consistent with the idea that retrieval is induced at each stage. Experiment 2 replicated the effect under more stringent conditions, and further showed that reductions in overconfidence are very similar for choice and for prompting retrieval directly, thus increasing the plausibility that increased retrieval drives the assessment method effect. Experiment 3 illustrated that an instruction to retrieve a large number of exemplars prior to a two-stage elicitation procedure reduced overconfidence even further. Nevertheless, several issues remain to be discussed.

Response Bias

Although the EBPJ gives an excellent account for the complete set of data, one other alternative proposal is that the choice and recall manipulations somehow influence people to merely report less confidence, rather than affecting their actual confidence. This kind of proposal has the flavor of a response bias account of overconfidence, whereby people are presumed to overstate their actual beliefs about the chances that events will occur. Overstatement of belief implies that people are unwilling to make consequential commitments as extreme as their reported probabilities imply. Although this issue was mentioned previously, its popularity as a proposed explanation for overconfidence demands further discussion.

Response bias accounts of overconfidence have been directly tested in the domain of general knowledge, but the tests have failed to produce supporting evidence for the idea. In one study, for example, Fischhoff, Slovic, and Lichtenstein (1977) instructed subjects extensively in the interpretations of various confidence scales, including the odds scale. They then offered subjects the opportunity to bet on their answers to general knowledge questions. Specifically, the subject would pay the experimenter a dollar for each item assigned at least 50:1 odds of being correct, but which was actually incorrect. For each of those same items, the subject would draw a poker chip from a bag containing 100 white chips, and 2 red chips (i.e. 50:1 odds of drawing a red chip). If a red chip was drawn, then the experimenter would pay the subject a dollar. The gamble would be in the subjects' favor if they were unbiased (because they would have reported even better than 50:1 odds for some of the items). Given typical levels of overconfidence, however, the gamble was very much against them. Yet, consistent with the idea that the subjects really did believe in the correctness of their answers at the stated confidence levels, they were almost always willing to take such unprofitable bets (also see Yates, Lee, & Bush, 1997). In the present study, participants in Experiments 2 and 3 were encouraged by the properness of the bonus system to be candid in their responses. Yet the choice and recall manipulations still led to reductions in overconfidence. Thus, it is especially unlikely that these manipulations encouraged people to report less confidence than they actually believed. Rather, those procedures appear to influence people's actual beliefs by encouraging them to retrieve more information than is "natural."

Choice Fundamentals and the Choice Model

Another issue concerns the question of whether or not choice behavior is fundamentally probabilistic or deterministic. Estes (1997) defined these possibilities as follows:

> Choice behavior is assumed to be *probabilistic* in the sense that, given full knowledge of the state of an individual's cognitive system at any time, we can only assign probabilities to his or her alternative choice responses. In contrast, we would speak of *deterministic* behavior if, given the same knowledge, we could predict with certainty the response that would be made (p. 326).

It should be clear from this definition that the NBPJ considered here assumes that choice behavior is probabilistic, whereas the EBPJ assumes that it is deterministic. According to the NBPJ, we can know for sure what an individual's subjective likelihood is in a particular situation, at least in principle. We cannot, however, know from that likelihood what choice the person will make. In contrast, the EBPJ implies that if we know the person's subjective likelihood at the time of decision, then we would know the person's choice for sure. The catch is that we cannot predict with certainty by any present means what an individual's subjective likelihood will be in a given situation.

A related question is whether or not the current study has any bearing on the validity of Luce's (1959) choice model. The gist of the choice model is that a respondent's probability of choosing a specific alternative among a set of possible alternatives is given by the ratio of the strength associated with that alternative to the sum of the strengths of all alternatives in the set. This rule can serve as the expression for choice probability in either of the two models under investigation here; only the definitions of "strength" differ. The interpretation for the adaptive network is straightforward. In the EBPJ, choice is predominately governed by the first exemplar retrieved. And as shown in Appendix B, probabilities of retrieval can be assigned to each exemplar that is stored in memory, based on the similarity of each to the present case. Hence in the choice model, the strength associated with an option, e.g. Trebitis, is essentially the total probability that an exemplar associated with Trebitis will be retrieved on the first cycle. Thus, the current study does not bear on the validity of the choice model, but only on the interpretation of its basic elements (see Estes, 1997, for some other interpretations).

Early Learning versus Expertise

The current study focused on the performance of individuals who gained fair amounts of experience in their judgment tasks. Nonetheless, there is convincing evidence that performance in these tasks changes when people acquire far more experience. For example, choice proportions tend to exhibit probability matching even well after a hundred trials, suggesting that matching is an effect that deserves special consideration. However, Edwards (1961) found that choices tended away from probability matching and towards probability maximization by the time subjects had accomplished a thousand trials (also see Ashby & Gott, 1988). Interestingly, choice proportions were not completely optimal by the end of a thousand trials, even in this simplest situation where there were no cues to attend to.

Of course, acquisition of great amounts of experience generally requires many distinct periods, and experiments employing very long trial lengths are typically conducted over several sessions. Hence, models intended to account for judgments of highly experienced individuals will need to incorporate assumptions regarding the nature of memory consolidation (e.g. Zola-Morgan & Squire, 1990). One possibility is that developing experts engage in undirected retrieval during periods of rest, and that responsiveness to retrieved exemplars is less attenuated than when there are external pressures to arrive at some decision. Such retrieval episodes could lead to deeper insights into the nature of the ecology, which are then stored in memory. Retrieval of these episodes that were formed off-line would then partially govern future classification behavior. A related, but distinct, possibility is that multiple memory systems are in operation, and that the system relied upon for judgment changes after sufficient experience. For example, McClelland, McNaughton, and O'Reilly (1995) postulated that the hippocampus forms an exemplar memory system, whereas the neocortex operates more like a connectionist learning system. The hippocampal system allows for rapid learning in new environments, whereas the neocortical system learns more general properties of the ecological structure, but at a much slower rate. Consolidation occurs because the slow-learning neocortical system is re-exposed to the hippocampal system's exemplars during off-line periods. Within this framework, judgments would tend to be based on the hippocampal system earlier in the learning process, but would be more often based on the neocortical system as much greater experience is achieved. An initial challenge for either of these approaches is determining more precisely the conditions under which judgments will be based on "pure" exemplars, and when they will be based on more consolidated memories. The present research suggests that, in either approach, responses should be taken as deterministically based on the information retrieved.

In the confidence judgment literature, experts in domains such as weather forecasting (Murphy & Winkler, 1977) and bridge playing (Keren, 1987) have been found to be essentially unbiased assessors of confidence. However, it is not clear whether that is due to sheer experience, or to special aspects of the tasks and training received. Further studies of learning and judgment in artificial ecologies need to be done to determine the conditions under which very highly experienced judges will be unbiased probability assessors.

Alternative Models of Likelihood Judgment

Although as mentioned in the introduction, models of category learning formed the theoretical point of departure for studying overconfidence in the specific task under investigation, the results have implications for several extant models of confidence judgment. For example, the theory of probabilistic mental models (PMM; Gigerenzer, Hoffrage, & Kleinbölting, 1991) was developed in the context of general knowledge questions, but might be even more descriptive of judgments about repeatable events. Indeed, the theory was tested using a special class of general knowledge items in which the alternatives were all highly similar. In PMM, cues are related to a target category via conditional probabilities which serve as cue validities (cf. Reed, 1972; Rosch & Mervis, 1975). The respondent chooses the alternative that is indicated by the cue with the highest validity. And the confidence value reported is equivalent to that validity. Because cue validities are assumed to equal their ecological counterparts, the model predicts that people are generally unbiased assessors of confidence. Overconfidence arises in standard general knowledge tasks because experimenters tend to select items that are unrepresentative of their respective populations. Specifically, the items tend to "trick" the respondent into thinking they are easier than they actually are.

Since the original statement of the PMM theory, considerable evidence has been levied against its explanation of overconfidence, typically by showing that overconfidence persists even with representative sampling of items (e.g. Brenner, Koehler, Liberman, & Tversky, 1996). And the current study shows that as well. In addition, the current research provides more direct evidence against PMM. Specifically, PMM assumes that, under conditions of representative sampling, people always choose the most likely option. Hence, it cannot account for probability matching. And, as shown in Experiment 1, assuming that choice is a probabilistic function of internal probability is not a feasable solution. That result provides significant problems for any theory, including PMM and the NBPJ, that assumes people operate primarily as "intuitive statisticians" (Peterson & Beach, 1967).

Another, very recent model for which the present results have implications was developed by Dougherty, Gettys, and Ogden (1999), based on earlier ideas by Hintzman (e.g. 1988). These authors' MINERVA-DM (MDM) relies heavily on memory representation to account for a variety of interesting phenomena in the judgment and decision literature. It also shares important commonalities with both the NBPJ and EBPJ described here. MDM assumes that experiences are stored as feature lists of individual episodes, generically called memory traces. In the ever continuing physician example, suppose the respondent is asked to judge the base rate of Trebitis. In this case, the similarity between the probe, "Trebitis," and all memory traces is computed, and then all of those similarities are summed to form a total level of activation, referred to as echo intensity. Echo intensity is assumed to be proportional to judged likelihood or frequency. Conditional likelihood judgments are formed by a two-stage process. Suppose in this case that the respondent is asked to judge the probability that a patient with a sore throat has Trebitis. First, echo intensity is evaluated based on all memory traces, using the symptom, sore throat, as the probe. Then, all of the traces that were sufficiently intense after the first round are activated again, this time by the disease, Trebitis. The respondent's conditional likelihood judgment is proportional to the echo intensity obtained from this latter search.

Overconfidence in judgment for repeatable events was among the phenomena investigated by Dougherty and his colleagues by simulation. In that simulation, the probability of a target event was based on the conditional echo intensity, as described above. Respondents were further assumed to always choose the disease for which the echo intensity was highest. The primary factors responsible for overconfidence in the model were the number of traces stored and the degrees to which those traces were intact. That is, since encoding is not perfect, each trace has "gaps" or is degraded to some degree. And poorly encoded experiences lead to overconfidence, because of resulting random error.

This model is very much like the EBPJ in its assumptions about memory representation. However, because it assumes that subjective likelihood is a function of all memory traces, MDM also behaves much like the NBPJ; except that it does not employ a probabilistic choice rule. Hence, it does not predict probability matching. And it does not anticipate the assessment method or retrieval instruction effects found in the current experiments. Nevertheless, it does possess more sophisticated assumptions regarding storage error than the present EBPJ. And, in simulations, overconfidence has been shown to follow from those premises. Experimental tests of the assumptions still need to be performed, however.

One other class of models that should be discussed are argument-based models, such as the argument recruitment model recently proposed by Yates, Lee, Shinotsuka, and Sieck (1999) to account for general knowledge overconfidence. The argument recruitment model draws on earlier ideas about the role of reasons in judgment (Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980), but proposes distinct assumptions about how the argument process leads to overconfidence, and its cross-cultural variations. The argument recruitment model proposes that, when confronted with a general knowledge question, a respondent (1) generates arguments that favor or oppose each of the alternatives, (2) assesses the extent to which the balance of the reasons favors one alternative or the other, (3) chooses the option indicated by the balance assessment, and (4) reports a probability judgment for the correctness of the chosen option according to the magnitude of the balance assessment. Two key principles concerning the recruitment process lead to overconfidence: (a) only a few arguments tend to be recruited, and (b) the recruitment process is biased toward the first argument that is generated.

Argument-based models have the potential to provide very general accounts of overconfidence, including overconfidence in tasks such as the physician's described here. However, a ubiquitous, tacitly held assumption in mapping these tasks onto argument-based models is that the cues displayed (e.g. symptoms) directly form the arguments. And some statistical measure of association between each cue and the target determines that argument's strength. The evidence suggests an alternative conception, in which the psychological association between cue and target is indirect. In this case, the cue prompts retrieval of an exemplar, and it is the exemplar that corresponds to an argument.

In the introduction of this article, we questioned whether the sources of overconfidence in general knowledge and repeatable events tasks were the same. The close correspondence between the mechanisms for overconfidence proposed in the exemplar retrieval and argument recruitment models imply that the sources might well be equivalent, or nearly so. And particular effects found in the current study suggest that as well. For example, the differences in overconfidence observed by method of elicitation parallel those found in studies of general knowledge (e.g. Ronis & Yates, 1987; Sniezek, Paese, & Switzer, 1990).

Although amount of recall has not been directly manipulated in general knowledge studies, there is indirect evidence that it does mediate overconfidence in those tasks. Specifically, Yates et al. (1999) had participants from three different cultures think out loud as they responded to general knowledge questions. They found that participants from Japan generated much longer protocols than did American or Chinese respondents, and also that the Japanese were the least overconfident. Although the correlation is compelling, the role of recall in general knowledge overconfidence needs to be directly tested in future research. One dissimilarity between the effects found in general knowledge and repeatable events tasks, however, is the amount of overconfidence observed. Specifically, in contrast to intuitions grounded in the "intuitive statistician" metaphor, the degree of overconfidence found here and in the Yates et al. (1998) study was far greater than that typically observed in studies of general knowledge. Nevertheless, the balance of evidence suggests that the processes underlying likelihood judgment in these distinct tasks are very similar, and that continued study of those processes will move us toward a universal account of overconfidence.

References

Anderson, N. H. (1973). Serial position curves in impression formation. *Journal of Experimental Psychology*, **97**, 8-12.

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

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, **37**, 93-110.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 33-53.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, **65**, 212-219.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, **106**, 180-209.

Dukas, R., & Real, L. A. (1993). Effects of recent experience on foraging decisions by bumble bees. *Oecologia*, **94**, 244-246.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, **62**, 385-394.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, **101**, 519-527.

Estes, W. K. (1957). Theory of learning with constant, variable or contingent probabilities of reinforcement. *Psychometrika*, **22**, 113-132.

Estes, W. K. (1964). Probability Learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 89-128). New York: Academic Press.

Estes, W. K. (1997). Some reflections on the role of the choice model in theories of categorization, identification, and learning. In A. A. J. Marley (Ed.), *Choice, decision, and measurement* (pp. 321-328). Mahwah, NJ: Erlbaum.

Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology*, **47**, 225-234.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 556-571.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 552-554.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1-67.

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

Gluck, M. A., & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General*, **119**, 105-109.

Gluck, M. A. (1992). Stimulus sampling and distributed representations in adaptive network theories of learning. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes, Vol. 1* (pp. 169-199). Hillsdale, NJ: Erlbaum.

Grant, D. A., Hake, H. W., & Hornseth, J. P. (1951). Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *Journal of Experimental Psychology*, **42**, 1-5.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41-58). New York: Seminar Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411-435.

Hagafors, R., & Brehmer, B. (1983). Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Decision Processes*, **31**, 223-232.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, **95**, 528-551.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, **24**, 1-55.

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, **25**, 294-301.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, **10**, 189-209.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, **39**, 98-114.

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, **10**, 279-285.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 107-118.

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

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, **20**, 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453-482). Chichester, England: Wiley.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**, 159-188.

Medin, D. L., & Florian, J. E. (1992). Abstraction and selective coding in exemplar-based models of categorization. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes, Vol. 2* (pp. 207-234). Hillsdale, NJ: Erlbaum.

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

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, **2**, 2-9.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psycholo;y*, **57**, 294-298.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes, Vol. 1* (pp. 149-167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 211-233.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, **104**, 266-300.

Paese, P. W., & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Organizational Behavior and Human Decision Processes*, **48**, 100-130.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, **43**, 87-131.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, **68**, 29-46.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, **253**, 980-986.

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

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

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, **40**, 193-218.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Russo, J. E., & Shoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, **33(2)**, Winter, 7-17.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *The Quarterly Journal of Experimental Psychology*, **42A**, 209-237.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 433-443.

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

Sieck, W. R., Quinn, C. N., & Schooler, J. W. (1999). Justification effects on the judgment of analogy. *Memory & Cognition*, **27**, 844-855.

Simon, H. A. (1956). A comparison of game theory and learning theory. *Psychometrika*, **21**, 267-272.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998) Alternative strategies of categorization. *Cognition*, **65**, 167-196.

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, **46**, 264-282.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, **65**, 117-137.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations, Vol. 1 (pp. 444-459). Cambridge, MA: Bradford Books/MIT Press.*

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, **67**, 201-221.

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers." *Psychological Bulletin*, **76**, 105-110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, **5**, 207-232.

Wilson, T. D., & LaFleur, S. J. (1995). Knowing what you'll do: Effects of analyzing reasons on self-prediction. *Journal of Personality & Social Psychology*, **68**, 21-35.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F., Lee, J.-W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and "reality." *Organizational Behavior and Human Decision Processes*, **70**, 87-94.

Yates, J. F., Lee, J.-W., Shinotsuka, H., & Sieck, W. R. (1999). *The argument recruitment model: Explaining general knowledge overconfidence and its cross-cultural variations.* Unpublished manuscript under editorial review, Department of Psychology, University of Michigan, Ann Arbor.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability accuracy: Beyond general-knowledge overconfidence? *Organizational Behavior and Human Decision Processes*, **74**, 89-117.

Zola-Morgan, S., & Squire, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science*, **250**, 288-290.

Appendix A

A Formal Instantiation of the Network-Based Probability Judgment Model

Assume that the respondent is considering a particular "patient's case," and that the patient is known to have exactly one of the two diseases "Trebitis" or "Philiosis." We denote by P_t the person's internal or covert probability that the patient actually has Trebitis, based on the symptoms displayed by the patient. According to the model,

$$P_t = \frac{1}{1+e^{-\sigma A}}, \tag{A1}$$

where A is the total activation favoring Trebitis, and σ is a scaling parameter. A is, in turn, given by

$$A = \sum_{i=1}^{m} w_i \alpha_i , \tag{A2}$$

where the α_i are indicator variables for each of m symptoms such that $\alpha_i = 1$ if the ith symptom is present, and $\alpha_i = 0$ if the ith symptom is absent. Further, w_i indicates the strength of the association that exists between the ith symptom and the disease category. That is, positive values of w_i tend to evoke the category Trebitis, and negative values evoke Philiosis; values near zero represent a lack of association between the symptom and diseases.

If the person is in a choice situation, he or she applies a probabilistic response rule to the internal probability, such that the probability of choosing Trebitis is equal to P_t. Reported confidence then depends on whether or not the choice was (subjectively) optimal or "perverse." If $P_t > .5$ and Trebitis is chosen, then the choice is optimal and P_t is the reported probability judgment that the choice is actually correct. Similarly, if $P_t < .5$ and Philiosis is chosen, then $1\text{-}P_t$ is the reported probability judgment that the choice is actually correct. However, if $P_t > .5$ and Philiosis is chosen, then the choice is subjectively non-optimal (perverse) and the reported confidence judgment (P_c) is given by

$$P_c = .5 + k(1\text{-}P_t), \tag{A3}$$

where k, with $0 \leq k \leq 1$, is a parameter indexing the person's willingness to give an extreme judgment after making a perverse choice. Similarly, $P_c = .5 + kP_t$, if $P_t < .5$ and Trebitis is chosen. Note that the reported probability that a perverse choice is correct increases linearly with the internal probability that that choice is actually correct.

If, instead of a choice, the task demands only a probability judgment that the patient has Trebitis, then P_t is reported. And 1-P_t is reported if the probability that the patient has Philiosis is directly requested.

Feedback regarding the disease category that actually obtained leads to learning via adjustments of the weights (association strengths), according to the Rescorla-Wagner (1972) learning rule. Imagine that the person has encountered j patients and has just received feedback concerning the jth patient's actual disease. According to the Rescorla-Wagner learning rule, the new association strength for symptom i is given by

$$w_{(j+1)i} = w_{ji} + \beta\alpha_i(T - A), \qquad \text{(A4)}$$

where β is a learning rate parameter that governs the amount by which the weight can change on a trial. T is an indicator variable such that $T = 1$ if the patient actually had Trebitis, and $T = -1$ if the patient had Philiosis. Note that the weights change more, the greater the disparity between the total activation and what actually occurred.

Figures A1 and A2 illustrate graphically the model's behavior under the ecology described in Experiment 1 as derived via Monte Carlo simulation (1000 simulated "subjects" per data point). The simulation assumed, for each level of β, a σ such that the mean P_t was about equivalent to the objective probabilities of Trebitis for that ecology to illustrate the model's performance when the internal probabilities are well calibrated.[8] The respective σs were 4.64, 3.23, 2.84, 2.71, 2.67 for βs ranging from .025 to .125, as shown in Figures A1 and A2. The simulation assumed that $k = .559$, which was estimated from the data of Experiment 1, via Equation A3. Observe that, for a range of plausible β values (Estes et al., 1989; Nosofsky et al., 1992), as anticipated, proportion correct is larger and overconfidence is lessened when probabilities are directly reported rather than when a choice is first made. Also note that there still exists overconfidence, even with direct reporting of the probability judgments. Subsequent analyses indicate that this bias is due to trial-by-trial variations in P_t which result from shifts in the weights during learning. Such effects are not further discussed here, because our primary focus is on the probabilistic response rule.

/ Insert Figure A1 about here /

/ Insert Figure A2 about here /

Appendix B

A Formal Instantiation of the Exemplar-Based Probability Judgment Model

Assume that the respondent is considering a particular "patient's case," and that the patient is known to have exactly one of the two diseases "Trebitis" or "Philiosis." We denote by $F_{t,N}$ the person's subjective probability that the patient actually has Trebitis, based on N retrieved exemplars. $F_{t,N}$ is the random variable given by

$$F_{t,N} = \gamma^N X_N + (1-\gamma^N)F_{t,N-1}\ , \tag{B1}$$

where

$F_{t,N-1}$ is the person's subjective probability of Trebitis, prior to retrieving the Nth case. $F_{t,0} = .5$ is the person's initial probability.

X_i are indicator variables for each of the outcomes in the sample of retrieved cases such that $X_i = 1$ if the ith "patient" in the sample had Trebitis, and $X_i = 0$ if the ith "patient" had Philiosis.

$\gamma (0 \leq \gamma \leq 1)$ represents the impact retrieved outcomes have on the person. Note that the impact of the most recently retrieved outcomes decreases with the number of retrieved exemplars, via the exponent.

N is a random variable representing the number of past cases that were retrieved. It follows a Poisson(μ) distribution, truncated so that $N > 0$.

The probability that a patient's case is retrieved from memory on each cycle is assumed to be governed by a modified version of the Medin and Schaffer (1978) retrieval rule. Imagine that the respondent has accumulated information on j-1 patients and is now facing the jth patient (the probe). According to the modified Medin-Schaffer retrieval rule, the probability that the kth patient (k = 1 to j-1) is retrieved is given by

$$P_k = \frac{s^{d_k + \frac{j-k}{j-1}}}{\sum_k s^{d_k + \frac{j-k}{j-1}}}\ , \tag{B2}$$

where d_k is the number of mismatching symptoms that exist between the *k*th patient and the probe (*j*th patient), and s ($0 < s \leq 1$) is a parameter that represents the similarity of mismatching values on a dimension. The relevant dimensions that influence retrieval are each of the symptoms and time. Equation B2 says that the probability that patient *k* will be retrieved decreases exponentially with each difference between patient *k*'s symptom profile and that of the probe, and the farther back is patient *k* in the set of accumulated patients (Nosofsky, 1984; Shepard, 1987). The parameter s can be thought of as measuring the degree to which respondents notice mismatching values. If $s = 1$, then the mismatching values are not noticed, so any patient is just as likely to be retrieved as any other, regardless of similarity to the probe. If $s = 0$, then any difference between patient *k* and the probe nullifies patient *k*'s chances of being retrieved. Because the Medin-Schaffer equation was modified here to include the time dimension, no stored patient is exactly like the probe, and so $s \neq 0$ in the present formulation.

Learning in the EBPJ occurs via the gradual accumulation of exemplars in memory. Specifically, as in some earlier versions (Estes et al., 1989; Nosofsky et al., 1992), the exemplar presented on each trial is stored in memory with some fixed probability, θ.

Figures B1 and B2 illustrate graphically the model's behavior under the ecology described in Experiment 1 as derived via Monte Carlo simulation (1000 simulated "subjects" per data point). The simulation assumed two representative values of s for each level of μ. The simulation further assumed $\gamma = .44$ as a representative value, based on responsiveness to single arguments in a think-aloud study (Yates et al., 1999). Finally, $\theta = 1$, because the primary focus here is on information retrieval, rather than storage. Observe that, as anticipated, there is little effect on the proportion correct and overconfidence decreases as the number of retrieved exemplars increases.

/ Insert Figure B1 about here /

/ Insert Figure B2 about here /

Footnotes

[1] Unless otherwise indicated, "overconfidence" will henceforth mean overconfidence in judgments about repeatable events.

[2] In fact, Dougherty, Gettys, and Ogden (1999) did so with Hintzman's MINERVA model. The present selections were made independently, and prior to publication, of that work. Nevertheless, the results of this study do have implications for the Dougherty et al. model, as is discussed in the conclusions chapter.

[3] The total activation is almost, but not exactly, on a probability scale. Specifically, values can fall outside of the 0-1 range. In the formal version of the model, a transform is used to correct this problem (see Appendix A).

[4] Overconfidence can also be traced to Step 5 (see Appendix A). However, the focus here is on that arising from Step 3.

[5] Boundary conditions on the probability matching phenomenon are discussed in the conclusions chapter.

[6] They also provide an alternative explanation of probability matching behavior. For example, suppose a respondent has accumulated 100 exemplars, on 70 of which one particular event occurred. If the judge bases a subsequent judgment on the retrieval of one past trial, then he has a 70% chance of saying that the event will occur, i.e., the judge probability matches.

[7] The probabilities from this model are essentially equivalent to corresponding probabilities derived via Bayes' theorem.

[8] The untransformed activation values do not always fall within the 0-1 range, so the purpose of Equation A1 is simply to make sure P_i does (Gluck & Bower, 1990).

Table 1

Description of ecology, Experiment 1

Symptom Ensemble	H	H,S	N,H	S	N,H,S	N	N,S
n	10	10	10	10	10	10	10
x	1	2	4	5	6	8	9

Note: n = number of times the ensemble was presented each session; x = number of times the target disease ("Trebitis") occurred, given the ensemble; the individual symptom abbreviations are N = runny nose, H = swollen hands, and S = sore throat.

Table 2

Regression Model Illustrating Influence of Recent Outcomes on Judgments, Experiment 1

	C50		NC100	
Model Term	Coef.	*t*	Coef.	*t*
Intercept	.45	10.64^{**}	.46	8.91^{**}
N	.11	5.76^{**}	.15	6.32^{**}
H	-.08	-4.48^{**}	-.01	-0.31
S	.08	4.66^{**}	.10	4.36^{**}
D_{-1}	-.03	-0.86	-.07	-1.69
C_{-1}	-.03	-2.39^{*}	-.03	-1.56
D_{-2}	-.06	-1.71	-.08	-1.95
C_{-2}	-.02	-1.34	-.04	-1.90
D_{-3}	-.02	-0.64	-.05	-1.03
C_{-3}	.00	-0.09	-.01	-0.69
$D_{-1} * C_{-1}$	.08	3.94^{**}	.07	3.08^{**}
$D_{-2} * C_{-2}$	.06	3.11^{**}	.07	2.78^{**}
$D_{-3} * C_{-3}$	.03	1.20	.05	1.75

$^{*}p < .05$; $^{**}p < .01$

Note: The abbreviations are N = runny nose, H = swollen hands, S = sore throat, D_{-i} = disease outcome of the *i*th past trial, and C_{-i} = number of common symptoms shared by the current and *i*th previous trial.

Table 3

Choice Proportions and Mean Probability Judgments, Experiment 1

Symptom Pattern	P(T)	Prop.(choose T)		P'(T)	
		C50	NC100	C50	NC100
H	.10	.21	.28	.31	.35
H, S	.20	.34	.38	.40	.42
N, H	.40	.31	.46	.40	.47
S	.50	.50	.49	.49	.50
N, H, S	.60	.67	.76	.62	.72
N	.80	.66	.65	.60	.62
N, S	.90	.75	.69	.67	.64

Note: P(T) = objective probability of the target disease ("Trebitis"), given the ensemble; Prop.(choose T) = proportion of times Trebitis was chosen; P'(T) = mean probability judgment of Trebitis. The individual symptom abbreviations are N = runny nose, H = swollen hands, and S = sore throat.

Table 4

Description of ecology, Experiment 2

Symptom Ensemble	N	N,S	N,H	N,H,S	H	H,S
n	15	15	15	15	15	15
x	1	4	7	8	11	14

Note: n = number of times the ensemble was presented each session; x = number of times the target disease ("Trebitis") occurred, given the ensemble; the individual symptom abbreviations are N = runny nose, H = swollen hands, and S = sore throat.

Table 5

Choice Proportions and Mean Probability Judgments, Experiment 2

	P(T)	Prop.(choose T)				P'(T)			
Symptom Pattern		Ctl.	Enc.	Rec.	C50	Ctl.	Enc.	Rec.	C50
N	.07	.19	.25	.24	.26	.26	.29	.31	.32
N, T	.27	.28	.38	.30	.32	.35	.40	.35	.36
N, H	.47	.47	.56	.46	.45	.50	.54	.47	.47
N, H, T	.53	.54	.49	.59	.48	.55	.48	.59	.47
H	.73	.56	.71	.65	.76	.60	.64	.63	.64
H, T	.93	.80	.78	.81	.77	.71	.69	.73	.68

Note: P(T) = objective probability of the target disease ("Trebitis"), given the ensemble; Prop.(choose T) = proportion of times Trebitis was chosen; P'(T) = mean probability judgment of Trebitis. The individual symptom abbreviations are N = runny nose, H = swollen hands, and S = sore throat.

Table 6

Choice Proportions and Mean Probability Judgments, Experiment 3

Symptom Pattern	P(T)	Prop.(choose T)		P'(T)	
		Recall	Control	Recall	Control
N	.07	.19	.29	.31	.32
N, S	.27	.16	.27	.37	.34
N, H	.47	.34	.48	.43	.46
N, H, S	.53	.27	.54	.41	.53
H	.73	.76	.71	.61	.61
H, S	.93	.84	.69	.70	.67

Note: P(T) = objective probability of the target disease ("Trebitis"), given the ensemble; Prop.(choose T) = proportion of times Trebitis was chosen; P'(T) = mean probability judgment of Trebitis. The individual symptom abbreviations are N = runny nose, H = swollen hands, and S = sore throat.

Figure Captions

Figure 1. Flowchart for Network-Based Probability Judgment model. *Notes:* P_T is the internal probability of the target (Trebitis), given encoded cues. For simplicity, P_T is assumed to be greater than .5 in this figure.

Figure 2. Flowchart for Exemplar-Based Probability Judgment model.

Figure 3. Flowchart for Network-Based Probability Judgment model in NC100 task. *Note:* P_T is the internal probability of the target (Trebitis), given the encoded cues.

Figure 4. Trial-by-trial procedure for each condition, Experiment 1. *Notes:* T is target category (Trebitis) and P is the alternative (Philiosis).

Figure 5. Bias (Overconfidence) by condition, Experiment 1. *Notes:* Bias = Mean probability judgment – Proportion correct; Bias > 0 indicates overconfidence. Error bars represent 1 standard error. Respondent chooses first, and then reports 50%-to-100% confidence that the choice is correct in C50 task. Respondent reports 0%-to-100% probability that a specified event will occur in NC100 task; the resulting probabilities are converted to corresponding choices and confidence values prior to calculating bias.

Figure 6. Trial-by-trial procedure for each condition, Experiment 2. *Notes:* T is target category (Trebitis) and P is the alternative (Philiosis).

Figure 7. Expected Bias (Overconfidence) by condition, Experiment 2. *Notes:* Expected Bias = Mean probability judgment – Expected proportion correct; Expected Bias > 0 indicates overconfidence. Error bars represent 1 standard error. Respondent reports 0%-to-100% probability that a specified event will occur in Control, Encode, and Recall tasks; the resulting probabilities are converted to corresponding choices and confidence values prior to calculating expected bias. Subjects in the Encode condition were instructed to carefully examine patient symptoms, prior to diagnosing each case. Subjects in the Recall condition were additionally instructed to bring to mind all similar patients seen previously. C50 subjects chose first, and then reported 50%-to-100% confidence that each choice is correct.

Figure 8. Trial-by-trial procedure for each condition, Experiment 3. *Notes:* T is target category (Trebitis) and P is the alternative (Philiosis).

Figure 9. Expected Bias (Overconfidence) by condition, Experiment 3. *Notes:* Expected Bias = Mean probability judgment – Expected proportion correct; Expected Bias > 0 indicates overconfidence. Error bars represent 1 standard error. Subjects in the Recall condition were instructed to recall similar patients, prior to diagnosing each case. Control subjects received no such instruction.

Figure A1. Proportion correct as a function of assessment method and learning rate. *Notes:* Respondent chooses first, and then reports 50%-to-100% confidence that the choice is correct in C50 task. Respondent reports 0%-to-100% probability that a specified event will occur in NC100 task; the resulting probabilities are converted to corresponding choices and confidence values prior to calculating proportion correct.

Figure A2. Bias (overconfidence) as a function of assessment method and learning rate. *Notes:* Bias = Mean probability judgment – Proportion correct; Bias > 0 indicates overconfidence.

Figure B1. Proportion correct as a function of amount of retrieval and similarity. *Notes:* The similarity parameter, s, indicates the extent to which the respondent notices differences between the present patient and past exemplars, with smaller values implying greater noticeability.

Figure B2. Bias (overconfidence) as a function of amount of retrieval and similarity. *Notes:* Bias = Mean probability judgment – Proportion correct; Bias > 0 indicates overconfidence.

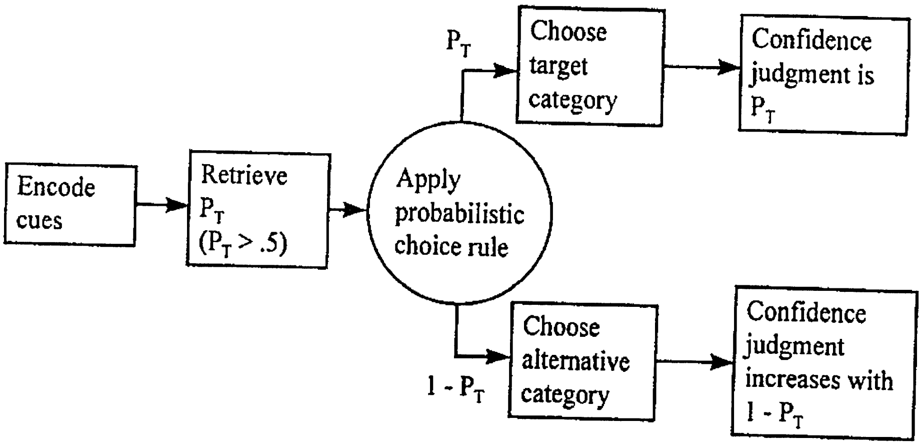
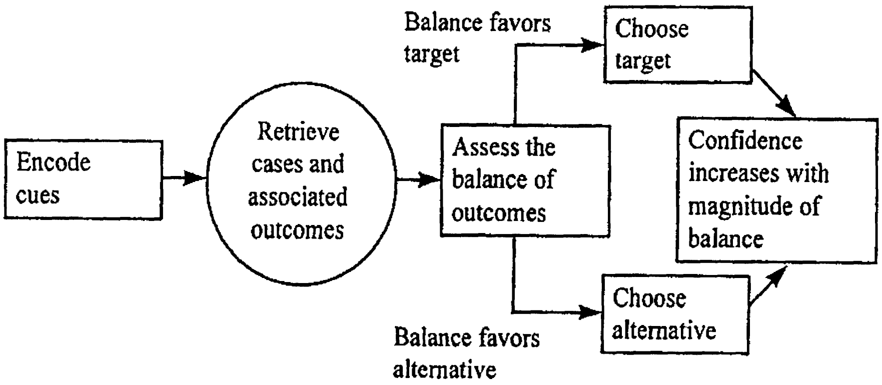

Encode cues
Retrieve P_T ($P_T > .5$)
Apply probabilistic choice rule
P_T
Choose target category
Confidence judgment is P_T
$1 - P_T$
Choose alternative category
Confidence judgment increases with $1 - P_T$

Figure 1


Encode cues
Retrieve cases and associated outcomes
Assess the balance of outcomes
Balance favors target
Choose target
Balance favors alternative
Choose alternative
Confidence increases with magnitude of balance

Key Assumption 1: Attenuated Responsiveness

Key Assumption 2: Abbreviated Retrieval

Figure 2

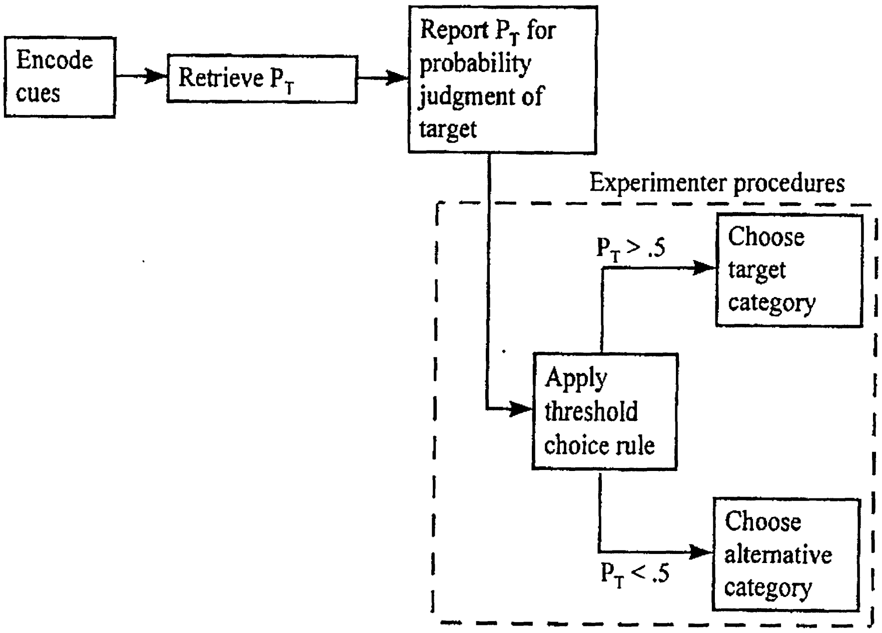

Encode cues
Retrieve P_T
Report P_T for probability judgment of target
Experimenter procedures
$P_T > .5$
Choose target category
Apply threshold choice rule
$P_T < .5$
Choose alternative category

Figure 3

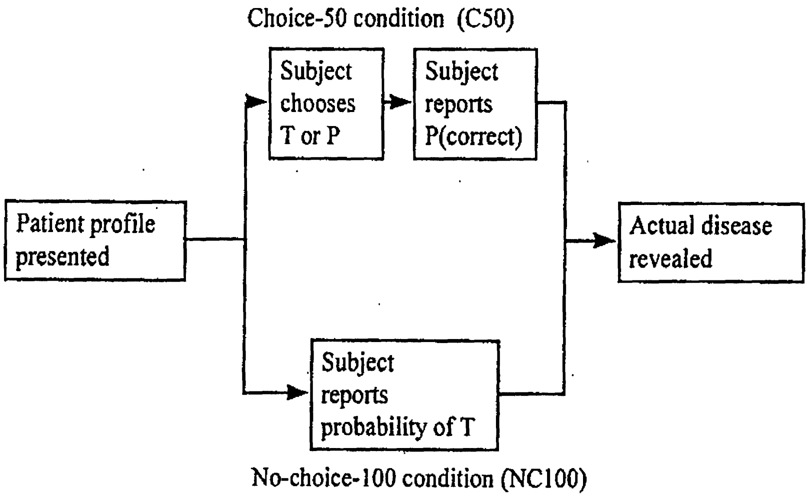

Choice-50 condition (C50)
Subject chooses T or P
Subject reports P(correct)
Patient profile presented
Actual disease revealed
Subject reports probability of T
No-choice-100 condition (NC100)

Figure 4

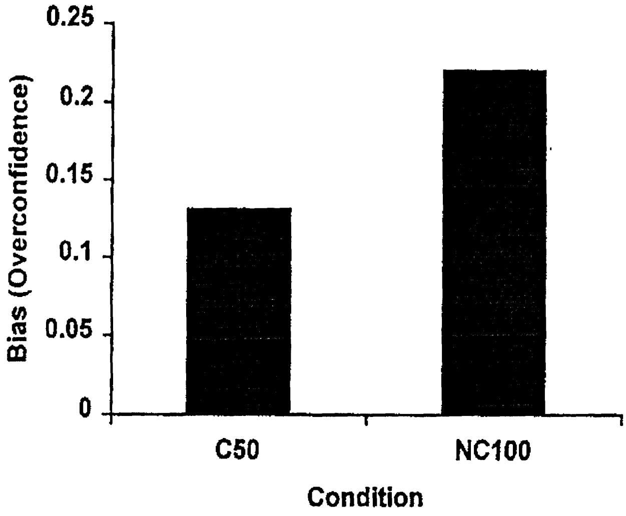


Figure 5

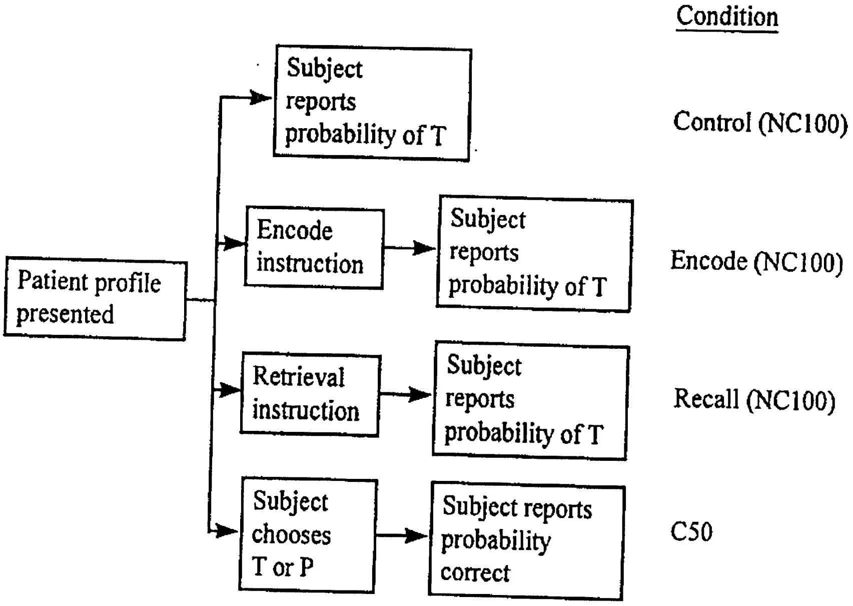

Condition
Patient profile presented
Subject reports probability of T
Control (NC100)
Encode instruction
Subject reports probability of T
Encode (NC100)
Retrieval instruction
Subject reports probability of T
Recall (NC100)
Subject chooses T or P
Subject reports probability correct
C50

Figure 6

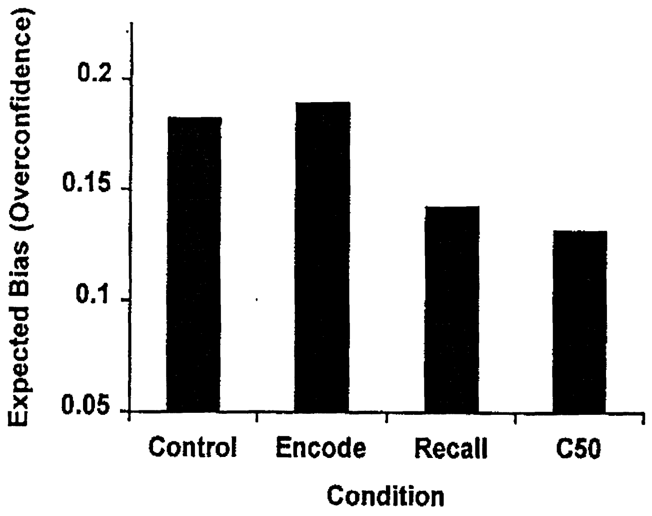

Expected Bias (Overconfidence)
0.2
0.15
0.1
0.05
Control
Encode
Recall
C50
Condition

Figure 7

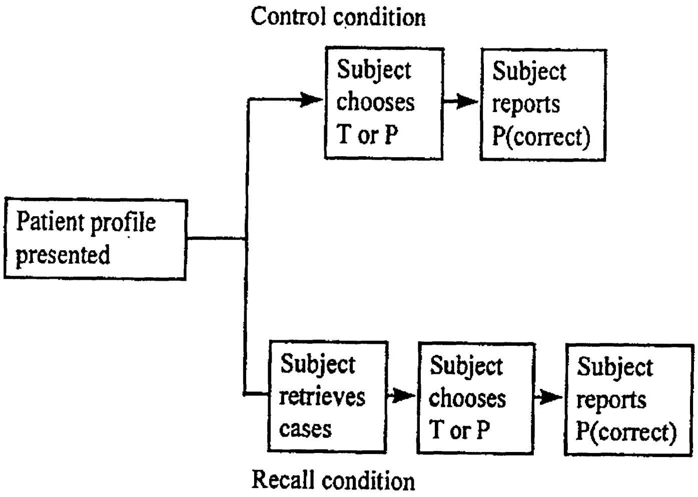

Control condition
Patient profile presented
Subject chooses T or P
Subject reports P(correct)
Subject retrieves cases
Subject chooses T or P
Subject reports P(correct)
Recall condition

Figure 8

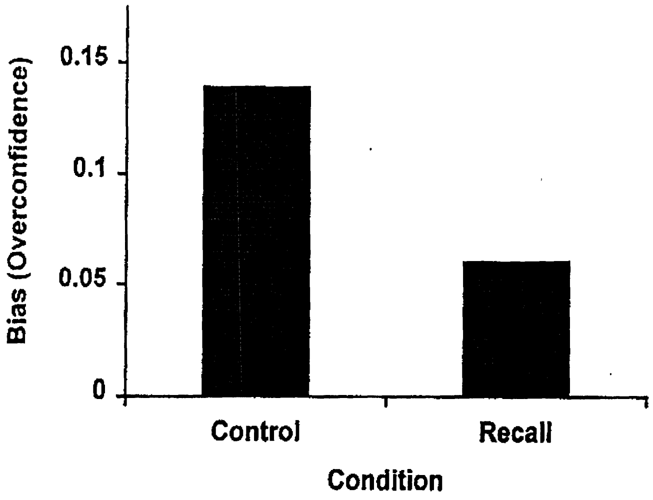


Figure 9

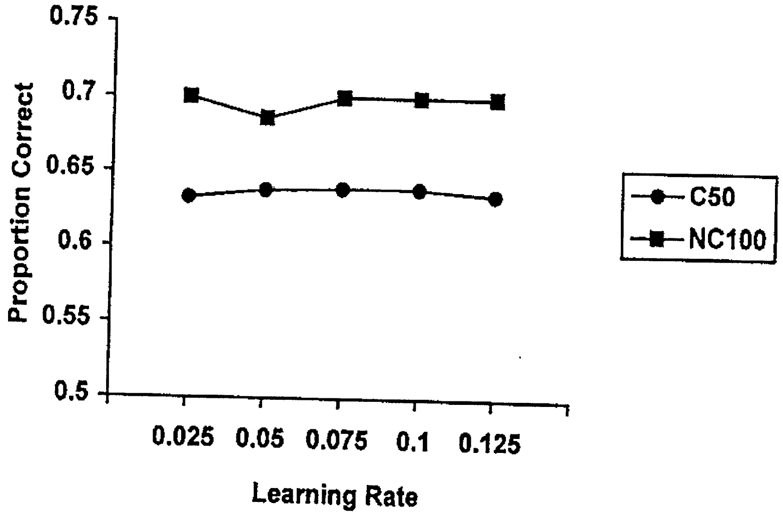

0.75
0.7
0.65
0.6
0.55
0.5
Proportion Correct
0.025
0.05
0.075
0.1
0.125
Learning Rate
C50
NC100

Figure A1

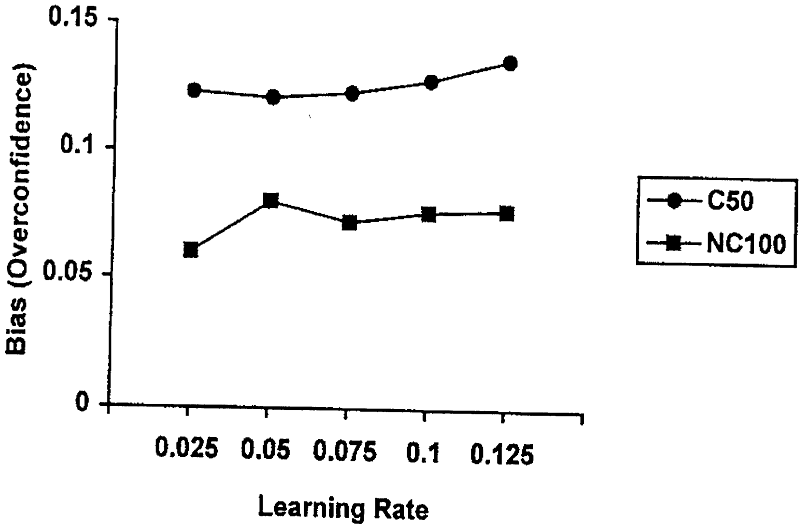

0.15
0.1
0.05
0
Bias (Overconfidence)
0.025
0.05
0.075
0.1
0.125
Learning Rate
—●— C50
—■— NC100

Figure A2

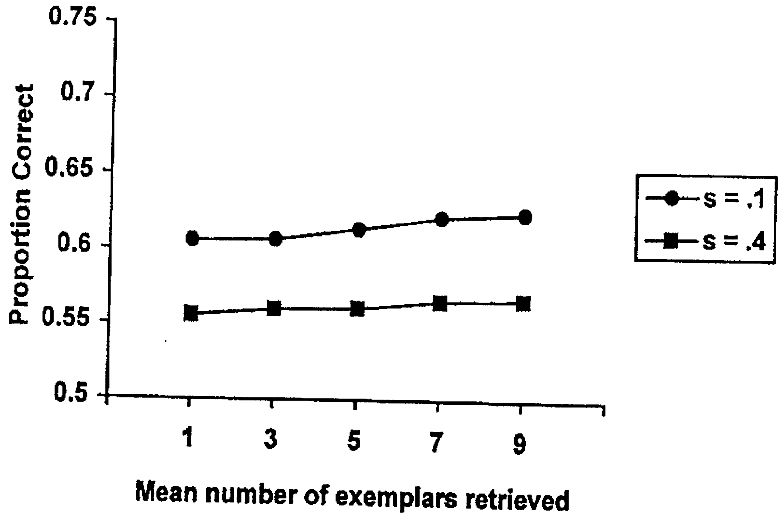

0.75
0.7
0.65
0.6
0.55
0.5
Proportion Correct
1
3
5
7
9
Mean number of exemplars retrieved
●s = .1
■s = .4

Figure B1

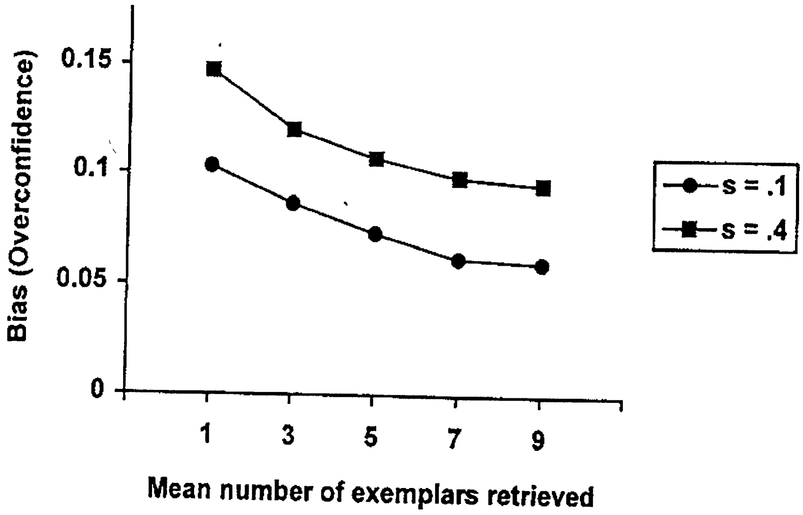

Bias (Overconfidence)
0.15
0.1
0.05
0
1
3
5
7
9
Mean number of exemplars retrieved
s = .1
s = .4

Figure B2