Journal of Experimental Psychology:
Learning, Memory, and Cognition
2001, Vol. 27, No. 4, 1003-1021

0278-7393/01/$5.00 DOI: 10.1037//0278-7393.27.4.1003

Overconfidence Effects in Category Learning: A Comparison of Connectionist and Exemplar Memory Models

Winston R. Sieck and J. Frank Yates
University of Michigan

Exemplar and connectionist models were compared on their ability to predict overconfidence effects in category learning data. In the standard task, participants learned to classify hypothetical patients with particular symptom patterns into disease categories and reported confidence judgments in the form of probabilities. The connectionist model asserts that classifications and confidence are based on the strength of learned associations between symptoms and diseases. The exemplar retrieval model (ERM) proposes that people learn by storing examples and that their judgments are often based on the first example they happen to retrieve. Experiments 1 and 2 established that overconfidence increases when the classification step of the process is bypassed. Experiments 2 and 3 showed that a direct instruction to retrieve many exemplars reduces overconfidence. Only the ERM predicted the major qualitative phenomena exhibited in these experiments.

Consider the following medical diagnosis problem:

Patient: K.M.
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 afflicting patient K.M. 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 K.M.'s condition, hence, providing the opportunity to improve his or her future diagnoses.

A past study that used a task like this one revealed an interesting overconfidence phenomenon. Specifically, participants were found to report average probability judgments that far exceeded the proportions of accurate diagnoses that were made (Yates, Lee, Shinotsuka, Patalano, & Sieck, 1998). For instance, in Experiment 1, American participants who had previous experience with 60 patients diagnosed another 60 patients. The mean 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 intense interest, and numerous proposed accounts are being vigorously debated (see, e.g., the recent reviews by Keren, 1997; McClelland & Bolger, 1994; Yates, Lee, Shinotsuka, & Sieck, 2001). However, 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 because performance in these distinct tasks is plausibly subserved by different memory systems. Specifically, the explicit factual information needed to respond to general knowledge questions should be located in a declarative memory storage system. In the current task, however, participants repeatedly classify patients into one of two disease categories on the basis of a fixed set of potential cues, and they receive feedback after each trial. So, performance in this task may well be mediated by a procedural memory system, as in cognitive skill learning or classical conditioning (Squire, 1986). Considerable evidence suggests that distinctions exist in the kind of information that is stored and retrieved in these separate systems, implying that overconfidence might occur for quite different reasons in the two tasks. Because memory representation and performance issues have been extensively investigated in studies

Winston R. Sieck and J. Frank Yates, Department of Psychology, Cognition, and Perception, University of Michigan.

This article is based on a doctoral dissertation submitted to the University of Michigan by Winston R. Sieck. This work was supported in part by National Science Foundation Grant SES-9911301, a University of Michigan School of Graduate Studies dissertation grant, and a University of Michigan Psychology Department thesis grant.

We thank Richard Gonzalez, Peter Lenk, David Meyer, and Richard Nisbett for invaluable criticisms, suggestions, and discussions regarding this research. This work also benefited from discussions with Matthew Jones, Shane Mueller, and Andrea Patalano. We are also grateful for helpful criticisms of previous versions of this article by Jerome Busemeyer, Shawn Curley, Paul Price, Thomas Wallsten, and two anonymous reviewers.

Correspondence concerning this article should be addressed to Winston R. Sieck, Department of Psychology, Cognition, and Perception, University of Michigan, 525 East University Avenue, Ann Arbor, Michigan 48109-1109. Electronic mail may be sent to sieck@umich.edu.

of category learning, ideas from that domain were expected to provide a useful theoretical point of departure for the current work.

Why is overconfidence in category learning tasks so significant? It is mainly 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. Any theory that endeavors to provide a universal account for overconfidence must explain its presence in category learning tasks as well as in those concerning general knowledge. And to the extent that specific overconfidence effects differ between these tasks, any potential unified theory will need to incorporate plausible memory and processing mechanisms that explain the differences. Unified theories are themselves critical for applications to more complicated, real-world judgment situations that might, for example, heavily engage both declarative and procedural memory systems.

This last point leads us to the practical consideration of the impact that overconfidence can have on decision quality and the very real costs that can be incurred. For instance, in the early 1970s, Royal Dutch/Shell noticed that newly hired geologists were overconfident in their predictions about the presence of oil or gas, despite having excellent credentials (Russo & Schoemaker, 1992). Specifically, only 1 or 2 wells in 10 produced when the geologists estimated a 40% chance of finding oil, thus costing the company enormous amounts of money and time.

Research Strategy

The goal of the present research was an improved understanding of overconfidence in category learning tasks. There is good reason to believe that many factors are capable of affecting overconfidence. The current focus, however, was on possible contributors that arise naturally from different memory systems. The theoretical approach was as follows. First, two well-known, relatively simple, models of classification learning that incorporate distinct memory systems 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 overconfidence in category learning[1] were then subjected to empirical investigation. Specifically, experiments were devised wherein the ordinal predictions of the models differed so as to obtain sharp tests between them.

The models adopted for extension were a simple connectionist network model (Gluck & Bower, 1988; Shanks, 1990) and an exemplar-memory model (Medin & Schaffer, 1978; Nosofsky, 1986). The simple network model essentially assumes that people learn relationships between cues and categories in a manner akin to classical conditioning. In contrast, the exemplar 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 as well (e.g., Anderson, 1991; Ashby, 1992; Hintzman, 1986; Kruschke, 1992).[2] The simple connectionist and exemplar models were chosen for initial comparison because their conceptions of memory representation and 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.

Activation Strength Model (ASM)

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. As the doctor gains experience by observing the symptom patterns of many patients and eventually finding out whether each had Trebitis or Philiosis, he or she comes to form associative bonds between each of the symptoms and the diseases. 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 or activation for one of the diseases over the other, say, Trebitis over Philiosis. These models assume that the doctor's diagnosis depends on this total degree of activation in a probabilistic fashion. That is, the probability that the doctor chooses Trebitis increases with the total degree of activation rather than the doctor always choosing Trebitis once some threshold is exceeded.

The present ASM extends these past ideas in a very natural way 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 to report a 50–100% confidence judgment. According to the ASM, the following steps occur.

Step 1: Symptom activation. Mental representations of the symptoms with which the patient presents become activated or "turned on." Symptom activation is denoted by α_i, 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.

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

[1] Unless otherwise indicated, *overconfidence* will henceforth mean overconfidence in category learning tasks.

[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 described in the General Discussion.

opposed to Philiosis (the assignment to Trebitis instead of Philiosis is arbitrary). That is, the total category activation is given by

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

where w_i indicates the strength of the association that exists between the ith symptom and the focal 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.

The total activation, A, is the person's internal or covert degree of belief or likelihood that the patient actually has Trebitis, based on the symptoms displayed by the patient. However, A generally falls within the range (-1 to 1), so the person's internal probability that the patient has Trebitis is given by

$$P_T = \frac{A + 1}{2}. \tag{2}$$

On those few occasions where A falls out of the (-1 to 1) bounds, P_T is assigned 0 or 1, respectively. As discussed below, P_T tends toward the ecological relative frequency of Trebitis as learning progresses.

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

Step 4: Probability judgment. Reported confidence that one's choice is correct is denoted by P_c and depends on whether the choice was (subjectively) optimal or "perverse." Suppose the respondent chose Trebitis. Then,

$$P_c = \begin{cases} P_T & P_T \geq .5 \\ .5 + kP_T & P_T < .5 \end{cases}, \tag{3}$$

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. One should note that the reported probability that a perverse choice is correct increases linearly with the internal probability that the choice is actually correct. If Philiosis is chosen, $P_P = 1 - P_T$, the probability for Philiosis would be substituted for P_T in Equation 3 (this rests on an assumption of binary complementarity, which has been shown to hold to a good approximation; e.g., Ariely et al., 2000).

After the classification process just described has ended, there is a learning step that occurs if feedback regarding the patient's actual disease is provided. According to the ASM, the learning occurs through Step 5.

Step 5: 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, through the Rescorla–Wagner learning rule (Rescorla & Wagner, 1972). 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

$$w_{(j+1)i} = w_{ji} + \beta\alpha_i(T - A). \tag{4}$$

In this expression, β is a learning rate parameter that governs the amount by which the weight (w) can change on a given trial. T is an indicator variable, such that $T = 1$ if the patient actually had Trebitis, and $T = -1$ if the patient had Philiosis. One should note that the greater the disparity between A and what actually occurred, T, the more the weights change.

The assumption of probabilistic choice in Step 3 has been critical to the past successes of network models in accounting for choice behavior, and it is key to the ASM's account of overconfidence. 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 nonoptimal 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 ASM'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 T actually occurs 70% of the time, people tend to predict the occurrence of event T 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). The Rescorla–Wagner learning rule leads to an activation level for the event T that tends toward the actual proportion of times that T occurs (Gluck & Bower, 1990). So, if people learn in a manner consistent with the Rescorla–Wagner equation, then there must be a rule that takes the internal probability as input and probabilistically returns a response in order to account for probability matching.[3]

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, of interest, in some early studies of probability learning, participants were instructed to "outguess the experimenter on each trial;" e.g., Estes & Straughan, 1954, p. 228). Analyses from game theory imply that a rational solution in these situations is to minimize the maximum difference between the actual payoff for a given set of strategies and the payoff that could be obtained if the adversary's strategy were anticipated (i.e., to minimax regret; Simon, 1956). When payoffs are symmetric, this implies that the optimal solution is to respond to the adversary's most likely action with the same probability that that action occurs. Hence, it could be that such a rule is induced from many previous experiences and then overgeneralized to all

[3] A problem with the probability matching prediction of these models is that choice proportions have been found to slightly overshoot objective probabilities greater than .5 and to slightly undershoot objective probabilities less than .5 after extremely long training periods. For example, Edwards (1961) found such effects by 1,000 trials of exposure to a single input pattern. Other investigators have attempted to deal with this problem by proposing probabilistic choice rules using a ratio of exponentials that explicitly contain response scaling parameters. Nevertheless, for learning into the hundreds of trials, predictions based on the Rescorla–Wagner rule do just as well without the extra response parameter as with it (see Gluck & Bower, 1988).

kinds of prediction tasks, such as the one under consideration here.[4] This argument is clearly speculative but may foster further needed discussion, as implied by the following quote from Estes (1997): "Throughout the history of research on learning and classification related to the choice model, the probabilistic conception seems to have been generally assumed, usually without discussion" (p. 326).

Exemplar Retrieval Model (ERM)

Exemplar models of category learning were originated by Medin and Schaffer (1978) and have been studied extensively since then (cf. Medin & Florian, 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. Specifically, Medin and Schaffer (1978) described the process as one in which classification is often based on the first exemplar retrieved. And the fundamental idea of classifications depending on one or a very few retrieved examples is also consistent with other current exemplar models (e.g. Nosofsky & Palmeri, 1997; Smith, Patalano, & Jonides, 1998).

The present ERM elaborates on the ideas of Medin and Schaffer (1978) in a very direct way 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 50–100% probability judgment that the classification was correct. According to the ERM, Steps 1–7 (below) occur.

Step 1: Encode cues. The patient's profile is examined.

Step 2: Exemplar retrieval. The respondent retrieves similar past instances, each of which indicates either Trebitis or Philiosis. Typically, only one exemplar is retrieved. The probability that a particular patient's case is retrieved from memory on each cycle is assumed to be governed by 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 Medin–Schaffer retrieval rule, the similarities along various symptom dimensions between the probe and previously stored patients are assumed to be combined in a multiplicative manner. Here, a single similarity parameter is used for all of the dimensions rather than using separate parameters for each dimension. According to this simplified version of the rule, the similarity between the jth and the kth patients ($k = 1$ to $j - 1$) is given by

$$sim(j, k) = s^{d_k}, \tag{5}$$

where d_k is the number of mismatching symptoms that exist between the kth patient and the probe (jth patient), and s ($0 \leq s \leq 1$) is a parameter that represents the similarity of mismatching values for each symptom. Equation 5 shows that the similarity between the previously stored patient k and the probe decreases exponentially with each difference between patient k's symptom profile and that of the probe (Nosofsky, 1984; Shepard, 1987). The parameter s can be thought of as measuring the degree to which respondents fail to notice mismatching values. If $s = 1$, then the mismatching values are not noticed, so that similarity to the probe is not influenced by that dimension. If $s = 0$, then any difference between patient k and the probe nullifies the overall similarity between those patients, regardless of how many other symptoms might match.

The probability that the kth patient is retrieved is given by

$$P(retrieve\ k \mid j) = \frac{sim(j, k)}{\sum_{k} sim(j, k)}, \tag{6}$$

and the total probability that any patient with the disease Trebitis is retrieved is

$$P(retrieve\ T \mid j) = \frac{\sum_{k \in T} sim(j, k)}{\sum_{k \in T} sim(j, k) + \sum_{k \in P} sim(j, k)}, \tag{7}$$

where T and P represent the disease categories Trebitis and Philiosis, respectively.

Step 3: Balance assessment. The respondent assesses the extent to which, on balance, the collection of exemplars retrieved in Step 2 supports either Trebitis or Philiosis. The balance assessment is represented by

$$S_N = \sum_{i=1}^{N} X_i, \tag{8}$$

where X_i indicates 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 $X_i = -1$ if the ith patient had Philiosis. N is a constant representing the number of past cases that were retrieved on each trial at the time of the choice response.

Step 4: Choice. The respondent chooses the disease favored in Step 3. That is, the respondent chooses Trebitis whenever $S_N > 0$ and chooses Philiosis whenever $S_N < 0$ (the response is random when $S_N = 0$). It is noted that the balance favors whatever exemplar was retrieved when $N = 1$, and so choice is completely governed by Equation 7 for that special case. Hence, it can be seen that the current model is a direct extension of the earlier formulation by Medin and Schaffer (1978).

Step 5: Exemplar retrieval. The prompt for a confidence judgment induces the respondent to engage in another retrieval episode. The probability that each patient retrieved during this episode has Trebitis is as given in Equation 7.

Step 6: Balance reassessment. The respondent reassesses the balance of the collection of exemplars obtained from Steps 2 and 5. The updated assessment is represented by

$$S_M = S_N + \sum_{i=N+1}^{M} X_i, \tag{9}$$

[4] A point against this argument is that respondents are often unable to generate sequences that pass statistical tests for randomness on demand (e.g., Bar-Hillel & Wagenaar, 1993). However, Rapoport and Budescu (1992; Budescu & Rapoport, 1994) found that respondents were more successful at generating random sequences in the context of actual two-person games than when explicitly instructed to do so.

where $M > N$ represents the total number of exemplars retrieved on the trial at the time of the confidence response.

Step 7: Probability judgment. The respondent's personal probability that the patient actually has Trebitis, based on the balance assessment, is denoted by $F_{T,M}$. $F_{T,M}$ is the random variable given by

$$F_{T,M} = \frac{\eta + M + S_M}{\eta + \phi + 2M}, \tag{10}$$

where $\eta/(\eta + \phi)$ represents the respondent's personal probability of Trebitis, prior to retrieving any past cases. Because, according to the model, this prior is not based on any real information, η and ϕ should be very small, and here we use $\gamma = \eta = \phi \in [0, 1]$, which form defensible Bayesian reference priors. The respondent's reported probability judgment is

$$F_{c,M} = \begin{cases} F_{T,M} & F_{T,M} \geq .5 \\ 1 - F_{T,M} & otherwise \end{cases}, \tag{11}$$

where $F_{c,M}$ is the reported probability that the patient actually has the indicated disease (i.e., is correct). Again, we are assuming binary complementarity, which seems to provide a good approximation (e.g., Ariely et al., 2000).

After the classification process has ended, there is a learning step that occurs if feedback regarding the patient's actual disease is provided. According to the ERM, the exemplar presented on each trial is deposited into a long-term store (LTS).

A key assumption of the ERM is that each retrieval episode generally produces only one exemplar, so that choice is based on a single exemplar and confidence is based on two exemplars (i.e., $N = 1, M = 2$). This abbreviated retrieval assumption is critical to the ERM's account of overconfidence.[5] The idea is that only a tiny fraction of a respondent's total knowledge store is incorporated into the judgment. Such a small sample provides a sizeable chance that the recruited evidence points in the wrong direction. However, because this sample is, according to the model, all that is taken into account, the respondent will be highly convinced in the appropriateness of the chosen alternative. It is noted that a second important assumption is that more information is retrieved with each judgment prompt. This implies that the confidence reported is not equivalent to the degree of confidence actually experienced at the time the choice was made.

These assumptions are further elaborated below. The first reason for expecting retrieval to be abbreviated 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. A reason to expect a second retrieval episode at the confidence prompt is that the new request influences people to reconsider the problem from a different angle. Research in memory retrieval suggests that repeated testing leads to changes in the retrieval cues used, so that recall is increased (e.g., Roediger & Payne, 1982). Also, judgment errors have tended to be reduced in judgment studies using within-participants rather than between-participants designs, suggesting that participants rethink problems when multiple requests are in place (e.g., Fischhoff, Slovic, & Lichtenstein, 1979).

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 they provided a 50–100% probability that the answer was in fact correct. In the second assessment method, items were initially circled (randomly) by the investigators, and then participants were asked to provide a probability, ranging from 0–100%, 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 much the same way as in the Ronis and Yates (1987) research. The ASM and ERM make opposing predictions about how overconfidence will differ in these two procedures, thus providing a direct test. Specifically, the ASM predicts that overconfidence will be reduced for the NC100 task, whereas the ERM predicts that it will be amplified. The reasoning behind these predictions is described as follows.

According to the ASM, 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 he or she must report a confidence level of at least 50%. Thus, such trials cause the respondent's judgments to exhibit overconfidence on the whole. But if the judge's task is to report a probability that the patient has Trebitis, without first indicating a choice, the respondent would render a judgment that directly reflects the experienced degree of activation (see Shanks, 1991). The researcher could then derive choices by applying a threshold response rule to the respondent's judgments (see Figure 1). This procedure eliminates the choice process, effectively bypassing the probabilistic response. Because overconfidence results largely from the probabilistic response rule, bypassing the stage in which the rule applies should substantially reduce overconfidence (see Appendix A for a more formal derivation).

A study by Neimark and Shuford (1959) lends empirical support for this proposal. In a standard probability learning task, those researchers had participants make predictions regarding a deck of cards that they turned up one after another. One group of participants predicted which of two letters would appear on each trial. Another group predicted similarly and also estimated the percent-

[5] It also provides 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 choice on the retrieval of one past trial, then he has a 70% chance of saying that the event will occur, that is, the judge probability matches.

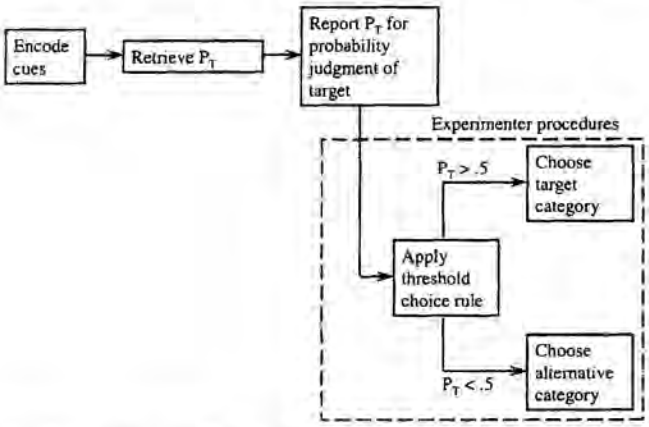

Figure 1. Flowchart for the activation strength model in the no-choice-100 task. P_T is the internal probability of the target (Trebitis), given the encoded cues.

age 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 Shuford's results are consistent with the ASM's assumption that people's choices will coincide with base rates probabilistically but that their estimates will match them directly, which implies that direct reporting of probabilities should reduce overconfidence, as previously argued.

The ERM makes just the opposite prediction. According to this model, judgments are based on small samples of exemplars that have been stored in memory, and more exemplars should be retrieved with more prompting. In the C50 task, there are two implicit demands for the participant 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, because participants are only prompted for a probability judgment. Hence, the probability judgments made by participants in the C50 task should be based on larger samples of exemplars than those made by participants in the NC100 task, so that greater overconfidence is found in the latter group (see Appendix B for a more formal derivation).

This proposal also rests on some empirical support. For example, Sniezek, Paese, and Switzer (1990) found greater overconfidence in a general knowledge task that used an NC100 procedure than in the same task that used a C50 procedure (also see Ronis & Yates, 1987). And they explained their findings by suggesting that confidence becomes more appropriate as the amount of cognitive processing increases.

Both accounts rest on some kind of experimental support, though neither of the past experiments is definitive with respect to the current issue of category learning overconfidence. The Sniezek et al. (1990) study directly assessed overconfidence, as we do here, but the probability learning task that Neimark and Shuford (1959) used arguably tapped the same memory systems as in the current research. Experiment 1 was conducted in order to resolve the conflict.

Method

Participants

Study participants were 86 undergraduates 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 description of the ecology for this article: (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; (c) 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.

Base rates and likelihoods. The base rate for both Trebitis and Philiosis was .50. Each of the symptoms had a base rate of .57. The likelihoods of N, H, and S given Trebitis were .77, .37, and .63, respectively.

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

$$p = \frac{e^Y}{1 + e^Y}, \tag{12}$$

where $p = P^*(T)$ was the model's probability judgment for the target event (in this case, Trebitis).[6] Y was the following linear combination of indicator variables, coded 1 and 0 for the presence and absence of the symptoms, respectively:

$$Y = -0.65 + 1.95\mathrm{N} - 1.63\mathrm{H} + .76\mathrm{S}. \tag{13}$$

The proportion of variance explained by the model was $R_L^2 = .24$.[7] 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. One was the standard C50 method, in which the participant was first asked to indicate the disease he or she felt each patient was most likely to have and then to provide a 50–100% probability that the patient actually had the

[6] In the present ecology, the probabilities from this model are essentially equivalent to corresponding probabilities derived through Bayes's theorem.

[7] There is no universally accepted definition of R^2 for logistic regression, so we report $R_L^2 = 1 - (D_\Omega/D_0)$, where D_Ω is the deviance (i.e., minus twice the maximized log-likelihood) for the tested model, and D_0 is the deviance for the null, or intercept-only, model (cf. Menard, 2000).

indicated disease. The second method was the NC100 assessment procedure, in which participants were asked to provide a probability, ranging from 0–100%, that each patient had the focal disease, Trebitis.

Procedure

The experiment was conducted entirely via computer. The program first introduced the scenario and initial instructions to the participant. 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 participant might have expected in a psychology experiment.

Participants were also given specific instructions concerning use of the probability scale. Participants 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.

Participants 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 participant (a) was presented with a new patient, who was identified by two initials and that patient's symptom profile; (b) indicated a probabilistic differential diagnosis according to either the C50 or NC100 procedure, as described above; and (c) received feedback about what was "eventually determined" to be the patient's actual condition. Participants made diagnoses for 70 patients during an initial block of trials and then returned to diagnose another 70 patients in a second block, 2 hr 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 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 nonfocal alternative. Choices were randomly selected for judgments of exactly 50%. Also, probability judgments that patients actually had the chosen diseases were derived from the C50 data by taking judgments "as is" when the focal disease was chosen and by subtracting judgments from 100% when the nonfocal alternative was selected. Overconfidence or underconfidence was indexed, as is the norm, through the following bias statistic:

$$\text{Bias} = \text{mean probability judgment} - \text{proportion correct}. \quad (14)$$

Positive values for bias indicate overconfidence, and negative values indicate underconfidence.

Model Simulations

The behavior of the two models under the ecology described in Experiment 1 was simulated through Monte Carlo methods (1,000 simulated participants per data point) in order to confirm the previously described predictions for the current experiment. Plausible ranges of the parameters were used to show that the predictions are not particularly sensitive to the parameter values chosen.

ASM. Figure 2 graphically illustrates the ASM's behavior under the ecology described in Experiment 1. The simulation assumed beta values ranging from .025–.125, which is a plausible range (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Nosofsky, Kruschke, & McKinley, 1992). The simulation further assumed three values for k that span its range of 0–1. One should observe that, 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

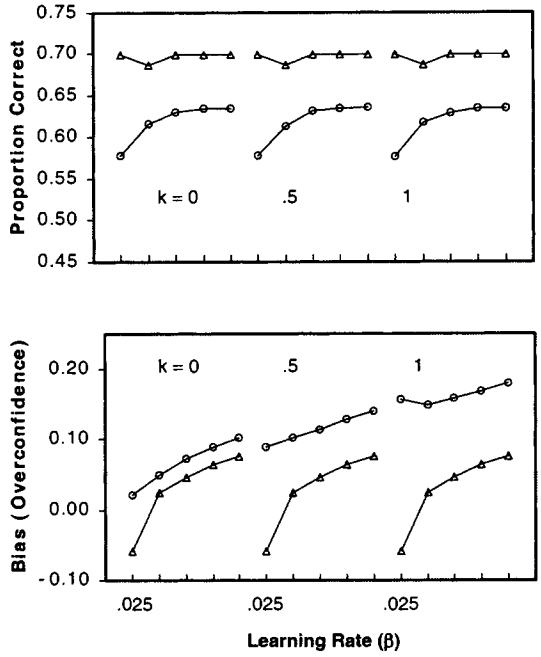

Figure 2. Simulated proportion correct and bias (overconfidence) according to the activation strength model, as a function of assessment method and learning rate in Experiment 1. Respondent chooses first and then reports 50–100% confidence that the choice is correct in the choice-50 (C50) task (C50 curve is denoted by ○). Respondent reports 0–100% probability that a specified event will occur in the no-choice-100 (NC100) task (NC100 curve is denoted by △); the resulting probabilities are converted to corresponding choices and confidence values.

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 learning. Such effects are not further discussed here because our primary focus is on the probabilistic response rule.

ERM. Figure 3 illustrates graphically the ERM's behavior under the ecology described in Experiment 1. The simulation assumed five values of s and three levels of γ, ranging from 0–1. One should observe that, as anticipated, proportion correct is unchanged and overconfidence increases when probabilities are directly reported rather than when a choice is first made.

Bias and Components

Table 1 shows the means and standard deviations of confidence, proportion correct, and bias for each condition. The C50-condition participants exhibited less confidence than the NC100-condition participants, but they achieved a slightly higher proportion of correct responses, $t(84) = -2.47, p = .015$, and $t(84) = 2.01, p = .048$, respectively. Participants in both conditions exhibited marked positive bias, that is, overconfidence. Also, the positive bias was much greater in the NC100 condition than in the C50 condition, $t(84) = 3.66$, $p = .0005$. Both of these results are opposite to the predictions of the ASM, but the confidence results support the ERM's predictions. The ERM, however, predicted that there would be no difference for proportion correct. This discrepancy is discussed below and is empirically addressed in Experiment 2.

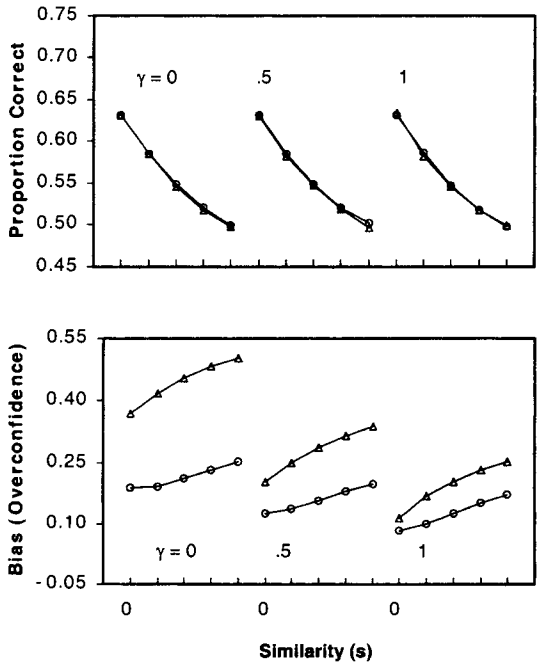

Figure 3. Simulated proportion correct and bias (overconfidence) from the exemplar retrieval model, as a function of assessment method and similarity in Experiment 1. Choice-50 curve is denoted by O; No-choice-100 curve is denoted by △.

Table 1
Means (and Standard Deviations) for Confidence, Proportion Correct, and Bias for Experiment 1

Condition	N	Confidence	Proportion correct	Bias
NC100	45	.79 (.108)	.57 (.078)	.22 (.128)
C50	41	.73 (.100)	.60 (.074)	.13 (.093)

Note. NC100 = no-choice-100; C50 = choice-50.

Judgment Given Symptom Pattern

Table 2 shows for each symptom pattern: (a) the objective probability of Trebitis for each symptom pattern, as defined under the ecology description; (b) the proportion of times Trebitis was chosen by participants 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 to 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 ASM. 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 may reflect that judgments are often inconsistent (cf. Erev, Wallsten, & Budescu, 1994). This result is addressed further in the General Discussion, but we note here that it is not diagnostic because both models anticipate such inconsistency, although by different mechanisms.

Summary

The findings of Experiment 1 provide some support for the ERM's account of category learning overconfidence. As proposed by that model, participants were less overconfident in the C50 task than in the NC100 task. And this finding was exactly opposite to the expectations of the ASM, thereby providing evidence against its account of overconfidence. Further evidence against the ASM's

Table 2
Choice Proportions and Mean Probability Judgments for Experiment 1

		Prop.(choose T)		P'(T)	
Symptom pattern	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; N = runny nose; H = swollen hands; S = sore throat; C50 = choice-50; NC100 = no-choice-100.

validity was that choices derived by a cutoff rule from the probability judgments given by participants 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.

Although the Experiment 1 overconfidence results were predicted by the ERM, there are difficulties with the interpretation. A limitation is that the C50-condition participants achieved a higher proportion of correct responses than did the NC100-condition participants. 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 was 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 were addressed in Experiment 2.

Experiment 2: Assessment Method and Recall

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

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

Another possibility was that participants in the NC100 condition felt it necessary to report more extreme probability judgments than did participants in the C50 condition, as described above. 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, participants 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 ERM, overconfidence should be reduced for this group, as in the C50 condition. However, this manipulation does not change the conversational demand because 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. This might happen because the second prompt induces people to reexamine the symptom profile. Such an attention effect might be expected according to the ERM through the similarity parameter, independent of the specific key assumption that abbreviated retrieval drives overconfidence. Any effect for the recall condition could also be interpreted as attentional in nature. In order to control for this possibility, we added an "encoding" condition wherein participants were instructed to pay close attention to the symptoms on each trial, in place of receiving the recall instruction or choice demand. Participants in this condition also made their judgments according to the NC100 procedure. This condition obviously controls for effects of attention.

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

Method

Participants

Study participants were 159 undergraduates 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.

Base rates and likelihoods. The base rate for both Trebitis and Philiosis was .50. The base rate for both symptoms N and H was .67, and the base rate for symptom S was .50. The likelihoods of N, H, and S given Trebitis were .44, .89, and .58, respectively.

Predictability. A logistic regression model describing Block 1 was created according to Equation 13. In this case, Y was the following linear combination of indicator variables for the presence and absence of the symptoms:

$$Y = -0.48 - 1.62\mathrm{N} + 1.62\mathrm{H} + .96\mathrm{S}. \quad (15)$$

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

Design

There were four between-participants conditions in Experiment 2: NC100 (control), NC100/encode (encode), NC100/recall (recall), and C50 (see Figure 4). All factors were introduced and no feedback was given in Block 2. The NC100 assessment method, described in Experiment 1, was used in the first three conditions. Participants in the control condition simply made judgments according to the NC100 procedure, as described in Experiment 1. Participants in the encode condition were presented with the following instruction for 3 s at the time the symptom profile was displayed, but prior to making their judgments: "Carefully examine this patient's

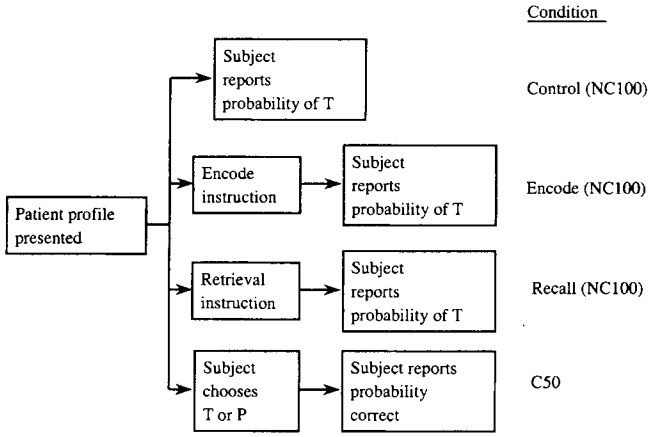

Figure 4. Trial-by-trial procedure for each condition in Experiment 2. T is the target category (Trebitis), and P is the alternative (Philiosis). NC100 = no-choice-100; C50 = choice-50.

symptom profile." Participants 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." Participants in the fourth condition made judgments according to the C50 procedure, as described in Experiment 1.

Procedure

The procedure was very similar to that of Experiment 1; the scenario and initial instructions were essentially identical. During Block 1, participants 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. Participants 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 participants of any procedure changes they would encounter in their diagnostic routine (i.e., specific condition instructions and corresponding probability judgment conventions). Participants 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 participants 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 participants with the four best accuracy scores would each receive a $20 bonus prize. This was intended to encourage effort and accuracy on the participants' part. The instructions emphasized that it was in the participant's best interest to be perfectly candid in reporting his or her true judgments to obtain their highest possible score. And the point was belabored by saying that it would hurt the participant's accuracy score to report any value other than his or her true belief. In Block 2, participants diagnosed another 30 patients according to one of the four conditions previously described. 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

Because feedback was not provided 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 "Trebitis" response would be correct over lots of possible sets of generated outcomes, and the objective conditional probabilities of Philiosis were similarly used to determine probabilities that each "Philiosis" response would be correct. These values were then averaged for each person to obtain the expected proportion of correct responses. Overconfidence or underconfidence were indexed through the following expected bias statistic:

$$E[\text{Bias}] = \text{mean probability judgment} - E[\text{proportion correct}]. \quad (16)$$

Model Simulations

The behavior of the two models was simulated through Monte Carlo methods under the ecology described in Experiment 2 in order to confirm the previously described predictions for the current experiment. The parameter settings were just as in Experiment 1 and again were intended to show that the model predictions are quite robust.

ASM. Figure 5 graphically illustrates the ASM's behavior under the ecology described in Experiment 2. One should observe that, as in the Experiment 1 ecology, proportion correct is larger and overconfidence is lessened when probabilities are directly reported rather than when a choice is first made. Also, it is noted that predictions for all of the NC100 groups are the same.

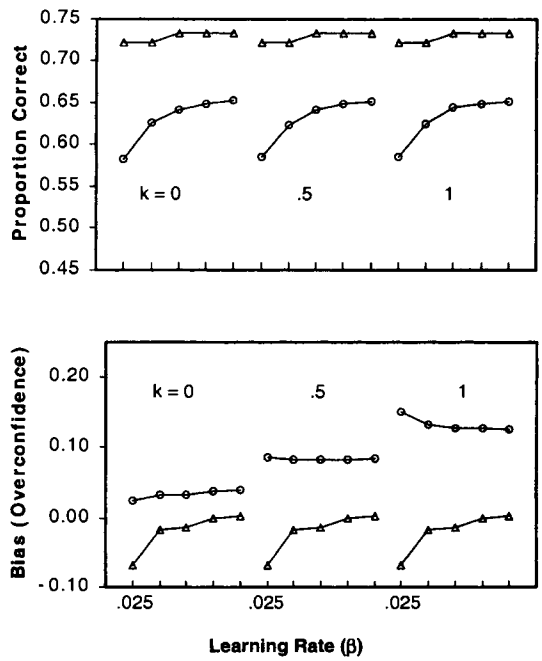

Figure 5. Simulated proportion correct and bias (overconfidence) from the activation strength model, as a function of assessment method and learning rate in Experiment 2. Choice-50 curve is denoted by O; No-choice-100 curve is denoted by △.

ERM. Figure 6 graphically illustrates the ERM's behavior under the ecology described in Experiment 2. One should observe that, as in the Experiment 1 ecology, proportion correct is unchanged and overconfidence is greater when probabilities are directly reported rather than when a choice is first made. Furthermore, the pattern of results is essentially equivalent when the choice prompt is replaced by a direct retrieval prompt. If the encode manipulation were to have any effect, it would be to reduce the value of the similarity parameter, *s*, so that proportion correct would be increased and overconfidence reduced.

Expected Bias and Components

Table 3 shows the means and standard deviations of confidence, proportion correct, and bias for each condition. The mean expected proportions of correct answers were essentially identical across conditions, $F(3, 155) = .074$, $p = .974$.

An analysis of variance (ANOVA) revealed a significant main effect for mean confidence, $F(3, 155) = 3.31$, $p = .022$. Confidence was nearly equivalent for participants in the control and encode conditions, $t(77) = 0.11$, $p = .915$. Those participants were more confident than participants in the recall condition, $t(117) = 2.14$, $p = .035$, and also more confident than participants in the C50 condition, $t(117) = 2.83$, $p = .005$. Finally, participants in the recall and C50 conditions were about equally confident, $t(77) = 0.69$, $p = .491$.

An ANOVA revealed a significant main effect for bias, $F(3, 155) = 2.67$, $p = .050$. Overconfidence was nearly equivalent for participants in the control and encode conditions, $t(77) = 0.26$, $p = .795$. Those participants were more overconfident than participants in the recall condition, $t(117) = 2.08$, $p = .040$, and also more overconfident than participants in the C50 condition, $t(117) = 2.30$, $p = .024$. Finally, participants in the recall and C50 conditions were about equally overconfident, $t(78) = 0.50$, $p = .618$. This pattern of results is just as anticipated by the ERM.

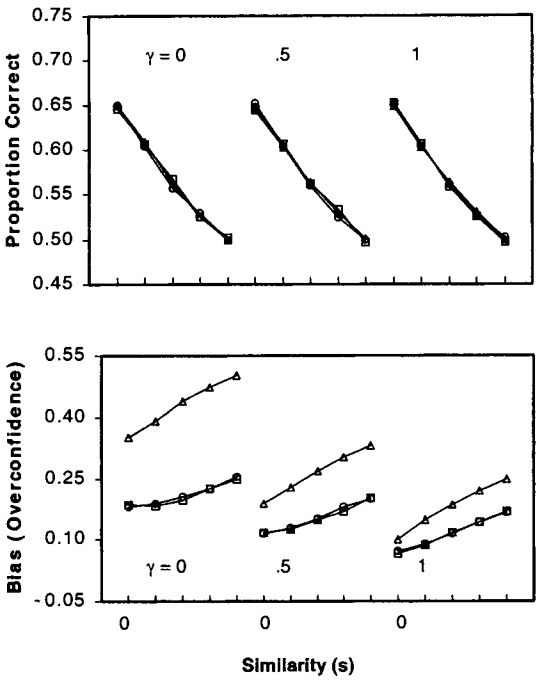

Figure 6. Simulated proportion correct and bias (overconfidence) from the exemplar retrieval model, as a function of assessment method, whether an extra retrieval prompt was included, and similarity, in Experiment 2. Choice-50 curve is denoted by ○; No-choice-100 curve is denoted by △; retrieval is denoted by □.

Table 3
Means (and Standard Deviations) for Confidence, Proportion Correct, and Bias for Experiment 2

Condition	*N*	Confidence	Proportion correct	Bias
Ctl./NC100	39	.79 (.093)	.61 (.096)	.18 (.125)
Enc./NC100	40	.78 (.102)	.60 (.109)	.19 (.122)
Rec./NC100	40	.75 (.086)	.61 (.089)	.14 (.066)
C50	40	.74 (.092)	.61 (.116)	.13 (.117)

Note. Ctl. = control; Enc. = encode; Rec. = recall; NC100 = no-choice-100; C50 = choice-50.

Judgment Given Symptom Pattern

Table 4 shows for each symptom pattern: (a) the objective probability of Trebitis; (b) the proportion of times Trebitis was chosen by participants in each condition; and (c) the mean probability judgments of Trebitis by condition. The group choice proportions matched the objective probabilities to a very rough approximation. As in Experiment 1, the mean probability judgments overshot the objective probability at the lowest values and undershot at the highest values.

Summary

In this experiment, the assessment method effect found in Experiment 1 was replicated under more stringent conditions. And the similarity in average overconfidence levels between the recall and C50 conditions increases the plausibility that recall drives that effect. Finally, no reduction in overconfidence was found in

Table 4
Choice Proportions and Mean Probability Judgments for Experiment 2

Symptom pattern	*P*(T)	Prop.(choose T)				*P'*(T)			
		Ctl.	Enc.	Rec.	C50	Ctl.	Enc.	Rec.	C50
N	.07	.19	.25	.24	.26	.26	.29	.31	.32
N, S	.27	.28	.38	.30	.32	.35	.40	.35	.36
N, H	.47	.47	.56	.46	.45	.50	.54	.47	.47
N, H, S	.53	.54	.49	.59	.48	.55	.48	.59	.47
H	.73	.56	.71	.65	.76	.60	.64	.63	.64
H, S	.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; Ctl. = control; Enc. = encode; Rec. = recall; N = runny nose; H = swollen hands; S = sore throat; C50 = choice-50.

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 (1995) 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. Because 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. Because individual records of past experiences simply do not exist according to the ASM, 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 ASM's explanation of overconfidence is accurate.

Experiment 3: Recall Prior to Choice

In Experiment 2, instructing participants to recall previously seen patients prior to reporting probability judgments was as effective at reducing overconfidence as an initial prompt for choice. The principal objectives of Experiment 3 were to replicate the recall effect and to determine whether such an instruction could reduce overconfidence over and above a demand to choose. The ERM predicts that overconfidence will tend to be further reduced as more and more exemplars are retrieved (see Appendix C for a derivation), so combined choice and recall demands are predicted to reduce overconfidence more than either demand alone. To determine this, we had control participants in this experiment follow the C50 procedure. Experimental participants performed the same task but were also instructed to recall as many similar patients as possible, prior to offering their diagnoses. In addition, including these multiple prompts to retrieve information should provide some idea of how effective mere recall can be as a debiasing strategy.

Method

Participants

Study participants were 56 undergraduates 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. Participants in the C50/recall (recall) condition were presented with the following instruction for 5 s 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 participants in the C50 (control) condition.

Procedure

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

Results and Discussion

Model Simulations

The Monte Carlo-simulated behavior of the ERM is shown in Figure 7 under conditions consistent with Experiment 3 to illustrate the effects of increasing the number of exemplars retrieved on proportion correct and bias. The parameter settings were generally like those of Experiment 1. The only difference is that the number of retrieved exemplars at the time of choice was varied to illustrate

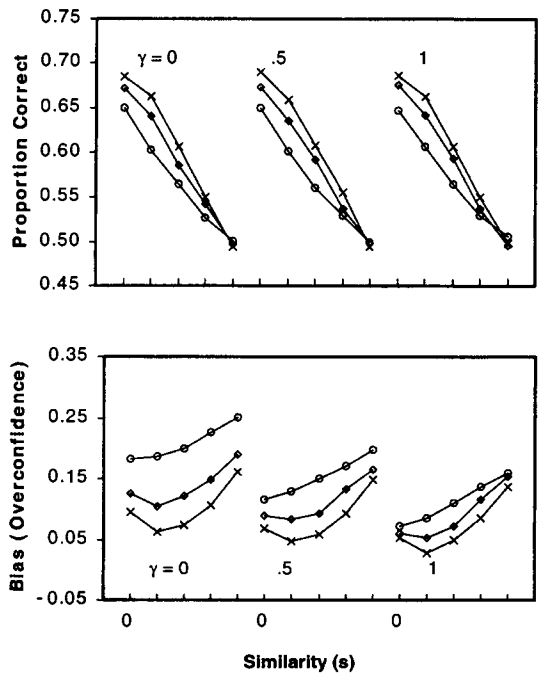

Figure 7. Simulated proportion correct and bias (overconfidence) from the exemplar retrieval model, as a function of number of exemplars retrieved and similarity in Experiment 3. One exemplar retrieved is denoted by ○; three exemplars retrieved is denoted by ◇; five retrieved is denoted by ×.

a range of predictions in the current experiment. In all cases, one more exemplar is retrieved at the time confidence is reported. One should observe that proportion correct increases somewhat and overconfidence decreases as the number of retrieved exemplars increases. No new simulations were conducted for the ASM because it does not predict any differences in the current experiment.

Expected Bias and Components

Table 5 shows that the participants who received the recall instruction exhibited slightly less confidence than participants who did not receive the instruction, although the difference is not statistically significant, $t(54) = -1.11$, $p = .137$ (one-tailed). Also, in the recall instruction participants achieved a slightly higher proportion correct, although the difference is only marginally significant, $t(54) = 1.52$, $p = .067$ (one-tailed). However, these small effects on the components combined so that participants in the recall condition exhibited less overconfidence than those in the control condition, $t(54) = -2.15$, $p = .036$. Yet, as can be seen in Table 5, overconfidence in the recall condition was still significantly positive, $t(27) = 2.23$, $p = .035$, suggesting that multiple retrieval demands do not readily eliminate overconfidence.

Judgment Given Symptom Pattern

Table 6 shows for each symptom pattern: (a) the objective probability of Trebitis; (b) the proportion of times Trebitis was chosen by participants in each condition; and (c) the mean probability judgments of Trebitis by condition. Choice proportions matched objective probabilities 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.

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. The pattern of results is, on the whole, well accounted for by the ERM. Further issues are addressed in the General Discussion.

Table 5
Means (and Standard Deviations) for Confidence, Proportion Correct, and Bias for Experiment 3

Condition	*N*	Confidence	Proportion correct	Bias
Ctl./C50	28	.73 (.109)	.59 (.124)	.14 (.131)
Rec./C50	28	.70 (.115)	.64 (.098)	.06 (.143)

Note. Ctl. = control; Rec. = recall; C50 = choice-50.

Table 6
Choice Proportions and Mean Probability Judgments for Experiment 3

		Prop.(choose T)		*P'*(T)	
Symptom pattern	*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; N = runny nose; H = swollen hands; S = sore throat.

General Discussion

The results of the three experiments reported here provide encouraging support for mechanisms of exemplar retrieval as contributors to overconfidence in category learning. An ASM 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 replicated the finding that retrieval reduces overconfidence and also found that a stronger retrieval manipulation did not eliminate overconfidence completely.

It is important to note that the failures of the ASM in this study do not imply that connectionist models cannot be used to represent information in an LTS. It may be possible, for example, to create a model with a connectionist representation of information in LTS that constructs exemplars at the time of retrieval. The constructed exemplars would then compose the retrieved information rather than activation signals, as in the ASM. Such a model would clearly share the critical assumption of the ERM that exemplars are retrieved and used in judgment. The ASM and ERM are parsimonious instantiations, at least conceptually, in that each proposes that stored and retrieved information take the same form (but see Estes, 1986, for an alternative view on model testing).

Limitations and Extensions

In this section, some limitations and possible extensions of the ERM are discussed.

Memory Strength

The ERM as presented in this article assumes that all presented exemplars are stored in memory equally. We view this as a simplification and suspect that it would be worthwhile to explore alternative formulations such as allowing exemplars to reside in memory with differing strengths. Such formulations would allow

the model to deal with factors such as the amount of time spent studying each case and the recency of presentation. Recency effects are particularly interesting because they may point toward another potential reason that retrieval is abbreviated. Specifically, abbreviated retrieval of recent events can be adaptive in autocorrelated ecologies (cf. Real, 1991). And recency effects were generally reported in the early probability-learning literature (e.g., Estes, 1957). Nosofsky et al. (1992) developed a recency-sensitive model by assuming that exemplar memory strength decays exponentially from the time of presentation. Busemeyer and Myung (1988) found that recency effects decreased as the number of presented cases increased, suggesting that an alternative formulation might be needed. This issue requires further investigation.

Variable Amounts of Retrieval

In the current rendition of the ERM, exactly one exemplar was assumed to be retrieved at each judgment prompt. This assumption was implicitly made by Medin and Schaffer (1978), and by Nosofsky (1986), and was explicitly made by Logan (1988) in his instance theory of automaticity. Nosofsky and Palmeri (1997) have recently developed an exemplar-based random walk model to predict categorization reaction times based on these earlier ideas, but which proposes that classifications rely on relatively small, random numbers of retrieved exemplars. This proposal needs to be further explored and tested in the context of overconfidence phenomena.

Retrieval Primacy

The ERM as described here tacitly assumes independence in exemplar retrieval. That is, the probability of retrieving an exemplar that points toward the target category on a given trial does not depend on the outcomes of exemplars previously retrieved on that trial. An alternative possibility is that retrieval is biased toward the categories of earlier retrieved exemplars. Some support for this idea comes from studies using paired associates. For example, Rundus (1973) found that presenting some of the items from a free-recall list to study participants interfered with their retrieval of the remaining items. This suggests that facts retrieved initially may serve as cues for further retrieval. The adoption of facts retrieved early in the process as cues for subsequent retrieval will, of course, foster dependencies between the facts ultimately retrieved. It is an open question as to whether such dependencies exist in the retrieval of exemplars in classification learning tasks.

Implications for 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, as described below.

Probabilistic Mental Models (PMM)

The theory of PMM (Gigerenzer, Hoffrage, & Kleinbölting, 1991) was developed in the context of general knowledge questions but might be even more descriptive of judgments in category learning tasks. In PMM, cues are related to a target category through conditional probabilities that 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). 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. As shown in Experiment 1, assuming that choice is a probabilistic function of internal probability is not a feasible solution. That result provides significant problems for any theory, including PMM and the ASM, that assumes people operate primarily as "intuitive statisticians" (Peterson & Beach, 1967).

Stochastic Models

Erev et al., (1994) defined a class of stochastic models of probability estimation and showed by simulation that random error in the judgment process was sufficient to produce overconfidence as revealed by calibration curves (where the judgment categories are plotted against the proportion correct given each judgment category). In the same simulations, the error also produced conservatism, that is, average probability judgments that overshoot objective probabilities below 50% and undershoot objective probabilities above 50%. Soll (1996), as well as Juslin, Olsson, and Björkman (1997), have extended the overconfidence results by showing that random error or inconsistency in the judgment process can also produce overconfidence as measured by the bias statistic used here. Although Erev, Wallsten, and Budescu's article did not attempt to specify the processes that lead to inconsistent judgments, it may well prove to be a watershed in the kinds of mechanisms that are brought to bear on overconfidence phenomena. For example, in the present study, we extended preexisting models of category learning to account for confidence judgments in very direct ways. Specifically, we did not build mechanisms for systematic overconfidence into the models. Overconfidence arises from both models principally because of stochastic error, although the specific mechanisms that lead to inconsistent judgments differ between the two models. Hence, the current findings support the conclusion that stochastic error is sufficient to produce overconfidence and expand that conclusion by exploring cognitive models that suggest points in the process where error might arise.

MINERVA-DM (MDM)

Another, very recent process model for which the present results have implications was developed by Dougherty, Gettys, and Ogden (1999), based on earlier ideas by Hintzman (1988). These authors' 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 ASM and ERM described here. MDM assumes that experiences are stored as feature lists of individual episodes ("traces") and that likelihood judgments involve similarity computations performed over all of the traces. Overconfidence in category learning was among the phenomena investigated by Dougherty et al. by simulation. In that simulation, respondents were assumed to choose deterministically. 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, because encoding is not perfect, each trace has gaps or is degraded to some degree. Poorly encoded experiences lead to overconfidence because of resulting random error.

This model is very much like the ERM 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 ASM, except that it does not use 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 sophisticated assumptions regarding storage error, and overconfidence has been shown to follow from those premises. Experimental tests of the assumptions still need to be performed, however.

Argument Recruitment Model (ARM)

One other class of models that should be discussed are argument-based models, such as the ARM recently proposed by Yates et al. (2001) to account for general knowledge overconfidence. The ARM 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 ARM proposes that when confronted with a general knowledge question, a respondent (a) generates arguments that favor or oppose each of the alternatives, (b) assesses the extent to which the balance of the reasons favors one alternative or the other, (c) chooses the option indicated by the balance assessment, and (d) 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 from 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 category learning tasks were the same. The close correspondence between the mechanisms for overconfidence proposed in the ERM and the ARM 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 according to method of elicitation parallel those found in studies of general knowledge (e.g., Ronis & Yates, 1987; Sniezek et al., 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. (2001) 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 category learning 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, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6,* 130–147.

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.

Bar-Hillel, M., & Wagenaar, W. A. (1993). The perception of randomness. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 369–393). Hillsdale, NJ: Erlbaum.

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.

Budescu, D. V., & Rapoport, A. (1994). Subjective randomization in one- and two-person games. *Journal of Behavioral Decision Making, 7,* 261–278.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3–11.

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.

Edwards, W. (1961). Probability learning in 1,000 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. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

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., 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.

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.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Decision Processes, 23,* 339–359.

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.

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.

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. (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.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

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., & 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.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *American Statistician, 54,* 17–24.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology, 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.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General, 121,* 352–363.

Real, L. A. (1991, August 30). 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 nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Roediger, H. L., III, & Payne, D. G. (1982). Hypermnesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 66–72.

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.

Rundus, D. (1973). Negative effects of using list items as recall cues. *Journal of Verbal Learning and Verbal Behavior, 12,* 43–50.

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

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 42*(A), 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, September 11). 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.

Squire, L. R. (1986, June 27). Mechanisms of memory. *Science, 232,* 1612–1619.

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.

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

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.

Yates, J. F., Lee, J.-W., Shinotsuka, H., & Sieck, W. R. (2001). *The argument recruitment model: Explaining general knowledge overconfidence and its cross-cultural variations.* Unpublished manuscript, University of Michigan, Ann Arbor.

(Appendixes follow)

Appendix A

Derivation of the Activation Strength Model's Choice Versus No-Choice Prediction

As described in the introduction to Experiment 1, overconfidence results in the standard C50 task primarily because of the probabilistic choice process. However, if the judge's task is to report a probability that the patient has Trebitis, without first indicating a choice, the respondent would directly report P_T. The researcher could then derive choices by applying a threshold response rule to the respondent's judgments (see Figure 3). This procedure eliminates the choice process, effectively bypassing the probabilistic response.

To show that no overconfidence is predicted for this NC100 procedure in the limit, suppose that input pattern j leads to T with probability π_j, and to the complement, T^c, with probability $1 - \pi_j$. Suppose next that the network settles so as to give output activation A_j to the input pattern j. Over all occurrences of pattern j, the expected mean squared error prediction, E_j, is

$$E_j = \pi_j(1 - A_j)^2 + (1 - \pi_j)(-1 - A_j)^2. \quad \text{(A1)}$$

The first term arises when T is the correct answer, and the second term arises when T^c is correct. Simplifying this equation and then taking the derivative of the expected error with respect to the activation and setting this equal to 0 gives

$$\frac{dE_j}{dA_j} = 2A_j + 2 - 4\pi_j = 0 \quad \text{(A2)}$$

and

$$\pi_j = \frac{A_j + 1}{2} = P_{T,j}. \quad \text{(A3)}$$

We thus find that the minimum squared error for pattern j occurs when the internal probability for that pattern is equivalent to the objective probability (see Gluck & Bower, 1990, for a very similar argument). So, no overconfidence is predicted for this procedure in the limit.

To show that overconfidence is expected for the C50 procedure, note that overconfidence or underconfidence is typically indexed through the following bias statistic: Bias = mean probability judgment − proportion correct. Also, suppose for simplicity that $.5 < \pi_j < 1$. Over all occurrences of the pattern j, the bias for that pattern is

$$\text{Bias}_j = [P_{T,j}(P_{T,j}) + (.5 + k(1 - P_{T,j}))(1 - P_{T,j})] - [\pi_j(P_{T,j}) + (1 - \pi_j)(1 - P_{T,j})]. \quad \text{(A4)}$$

Next, since a parameter-free derivation is desired, we let k take on its minimum value of 0 without loss of generality. Substituting π_j for $P_{T,j}$ by Equation A3, Equation A4 then simplifies to

$$\text{Bias}_j = \pi_j^2 + .5(1 - \pi_j) - \pi_j^2 - (1 - \pi_j)^2 = (\pi_j - .5)(1 - \pi_j). \quad \text{(A5)}$$

And because both terms on the right-hand side of Equation A5 are positive, the bias is positive, thus indicating overconfidence. Although this argument is compelling, it relies on the ASM's assumption that people report their true beliefs; that is, their probability judgments directly reflect experienced activation strength. An alternative, less restrictive, formulation might suppose that some response bias is in operation. To capture this, say people report $G_{T,j} = g(P_{T,j})$, where $g(\cdot)$ is a monotonic, nondecreasing function of internal probability. Then, recalling that we are dealing with $.5 < \pi_j < 1$ for convenience, the bias for the NC100 task is $G_{T,j} - \pi_j$. Following the same derivation as above, we find that, for the C50 task,

$$\text{Bias}_j = (G_{T,j} - \pi_j)\pi_j + (\pi_j - .5)(1 - \pi_j). \quad \text{(A6)}$$

Comparing these two expressions, it is clear that overconfidence will be greater in the C50 task, so long as

$$G_{T,j} - \pi_j < \pi_j - .5 \quad \text{(A7)}$$

and

$$G_{T,j} < 2\pi_j - .5. \quad \text{(A8)}$$

Examination of Equation A8 reveals that greater overconfidence will be found in the C50 task, irrespective of any model parameter value settings, when $\pi_j > .75$. It is also apparent that for most ecological probabilities, any response bias must be quite large for the predictions to fail (e.g., the response bias must be greater than .15 to obtain a reversal in predictions when $\pi_j = .65$). This derivation is somewhat limited by the assumptions needed to make it manageable, such as infinite learning and repeated exposure to only a single pattern. We show by simulation that the predictions hold under more realistic conditions.

Appendix B

Derivation of the Exemplar Retrieval Model's Choice Versus No-Choice Prediction

According to the ERM, overconfidence results in the standard C50 task primarily because of the abbreviated retrieval process. Retrieval abbreviation is exacerbated in the NC100 task because it contains only one prompt for information, whereas the C50 task has two. To show this, as in the setup for the network model derivations, suppose that input pattern j leads to T with probability π_j $(.5 < \pi_j < 1)$ and to the complement, T^c, with probability $1 - \pi_j$. To make the derivations more tractable, suppose that the respondent is exposed only to pattern j so that $\pi = \pi_j$, then $d_k = 0$ for all presented exemplars and the retrieval rule (Equation 11) simplifies to

$$P(\text{retrieve } T \mid j) = \frac{\sum_{k \in T} 1}{\sum_{k \in T} 1 + \sum_{k \in P} 1} = \frac{n_T}{n_T + n_P} = \pi. \quad \text{(B1)}$$

As in Appendix A, overconfidence is measured by the difference between the mean probability judgment and the proportion of correct responses. Because choice is based on the retrieval of one exemplar in the C50 task, the proportion correct should be equivalent for the C50 and NC100 tasks. Hence, it is sufficient to show that the mean probability judgment is lower when based on two exemplars rather than just one. The mean probability judgment based on one exemplar is

$$\text{E}[F_{c,M=1}] = \pi\left(\frac{\gamma + M + 1}{2(\gamma + M)}\right) + (1 - \pi)\left(\frac{\gamma + M - (-1)}{2(\gamma + M)}\right) = \frac{\gamma + 2}{2(\gamma + 1)}, \quad \text{(B2)}$$

and the mean probability judgment based on two exemplars is

$$\mathrm{E}[F_{c,M=2}] = \pi^2\left(\frac{\gamma + M + 2}{2(\gamma + M)}\right) + 2\pi(1 - \pi)\left(\frac{\gamma + M + 0}{2(\gamma + M)}\right) + (1 - \pi)^2\left(\frac{\gamma + M - (-2)}{2(\gamma + M)}\right), \quad \text{(B3)}$$

which reduces to

$$\mathrm{E}[F_{c,M=2}] = -2\pi(1 - \pi)\left(\frac{1}{\gamma + 2}\right) + \frac{\gamma + 4}{2(\gamma + 2)}. \quad \text{(B4)}$$

Comparing Equations B2 and B4, we find that the mean probability judgment based on two exemplars will be less than the mean judgment based on one exemplar when

$$-2\pi(1 - \pi) < \left[\frac{\gamma + 2}{2(\gamma + 1)} - \frac{\gamma + 4}{2(\gamma + 2)}\right](\gamma + 2). \quad \text{(B5)}$$

Solving for γ yields the inequality

$$\frac{4\pi(1 - \pi)}{1 - 4\pi(1 - \pi)} > \gamma. \quad \text{(B6)}$$

According to Equation B6, in order for overconfidence in the C50 condition to be lower than in the NC100 condition, γ should be small and the objective probability, π, should be extreme. However, the C50 versus NC100 result holds over the values of γ so long as π is not too extreme (i.e., for all $.15 < \pi < .85$). Hence, the result is quite robust. However, this derivation is limited by the assumption of repeated exposure to one pattern, so we also show that the predictions hold under more realistic conditions by simulation.

Appendix C

Derivation of the ERM's Increased Retrieval Prediction

According to the ERM, overconfidence results because of the abbreviated retrieval process. So, overconfidence ought to be diminished by increasing the amount of information retrieved. To show this, we assume that the respondent is exposed to only one pattern, as in Appendix B, so that Equation B1 holds. The mean probability judgment and proportion correct are given, respectively, by

$$\bar{f} = \mathrm{E}[F_{T,M}|S_N \geq 0]P(S_N \geq 0) + \mathrm{E}[1 - F_{T,M}|S_N < 0]P(S_N < 0) \quad \text{(C1)}$$

and

$$\bar{c} = \pi P(S_N > 0) + (1 - \pi)P(S_N < 0) + .5P(S_N = 0), \quad \text{(C2)}$$

where $M = N + 1$. The probability that $S_N > 0$ increases toward one as N increases. To see this, one should first note that $S_N > 0$ if $S_N/N > 0$, and by the central limit theorem, S_N/N converges to the normal pdf with mean $2\pi - 1$ and variance $4\pi(1 - \pi)/N$. So, $P(S_N > 0) \to 1$ as $N \to \infty$. By applying this fact to the difference between Equations C1 and C2, and noting that the mean of the personal probability is

$$\mathrm{E}[F_{T,M}] = \frac{\gamma + M + M(2\pi - 1)}{2\gamma + 2M} = \frac{\gamma + (2\pi)M}{2\gamma + 2M}, \quad \text{(C3)}$$

we have

$$\underset{N\to\infty}{Bias} = \mathrm{E}[F_{T,M}] - \pi = \mathrm{E}[F_{T,N+1}] - \pi = \frac{2\pi}{2} - \pi = 0. \quad \text{(C4)}$$

That is, bias decreases to 0 as N gets very large. The fanciful assumptions, such as infinite retrieval and repeated exposure to only one pattern, obviously limit the conclusions. Thus, we also use simulation to show that overconfidence decreases with retrieval under more realistic assumptions.

Received January 6, 2000
Revision received August 16, 2000
Accepted January 17, 2001 ■