Journal of Behavioral Decision Making
J. Behav. Dec. Making, **16**: 127–145 (2003)
Published online 19 February 2003 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.438

Effects of Choice and Relative Frequency Elicitation on Overconfidence: Further Tests of an Exemplar-retrieval Model

WINSTON R. SIECK*
Ohio State University, USA

ABSTRACT

An experiment is reported in which participants rendered judgments regarding the disease states of hypothetical patients. Participants either reported likelihoods that patients had the target disease (no choice), or classified patients into disease categories and then reported likelihoods that their classifications were correct (choice included). Also, participants' likelihood judgments were made in response to either a probability probe question, or a relative frequency probe. Two distinct exemplar-memory models were compared on their ability to predict overconfidence under these procedures. Both propose that people learn and judge by storing and retrieving examples. The exemplar retrieval model (ERM) proposes that amount of retrieval drives choice inclusion and likelihood probe effects. The alternative model assumes that response error mediates choice inclusion effects. Choice inclusion and the relative frequency probe reduced overconfidence, but the combined effects were subadditive. Only the ERM predicted this pattern, and it further provided good quantitative fits to these results. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS confidence; subjective probability; calibration; classification learning

An emergency room nurse at a major hospital makes numerous triage decisions throughout the day. Patients are classified as to whether they need immediate treatment or not based on the symptoms they present. Sometimes the nurse is quite confident that a correct decision has been made, whereas other times the situation seems less clear. Furthermore, the nurse's experience with past patients will determine, in some way, how new patients are classified and with what degree of confidence. Likewise, a financial analyst studies the profiles of many companies, and based on key characteristics of those profiles, renders judgments as to whether each will exceed or fall short of expected earnings for the quarter. Admissions committee members examine prospective students' portfolios and predict whether they will be successful in the program. On what basis are such judgments made? A clue as to how they are at least sometimes accomplished can be gleaned from

*Correspondence to: Winston R. Sieck, Department of Psychology, Ohio State University, 1827 Neil Avenue, Columbus, OH 43210-1222, USA. E-mail: sieck.3@osu.edu

Contract/grant sponsor: National Science Foundation.
Contract/grant numbers: SES-9911301; SBR-0196200.

Camerer and Johnson's (1991) discussion of a verbal protocol produced by the chair of a hospital's admissions committee: 'Seeing an applicant from Wayne State who had very high board scores, the doctor recalled a promising applicant from the same school who had perfect board scores. Unfortunately, after being admitted, the prior aspirant had done poorly and left the program. The physician recalled this case and applied it to the new one: ''We have to be quite careful with people from Wayne State with very high board scores . . . We have had problems in the past'' ' (p. 209).

In the categorization literature, exemplar-based models that characterize the classification process as essentially like that of the physician above have enjoyed considerable success over the last couple of decades. Recently, several studies have examined confidence in such classification tasks (Dougherty, 2001; McKenzie, 1998; Sieck & Yates, 2001; Yates et al., 1998), and there have been some theoretical developments concerning confidence that incorporate exemplar representations. For example, Sieck and Yates' (2001) exemplar retrieval model (ERM) extends Medin and Schaffer's (1978) context theory of classification learning to account for confidence judgments in addition to categorization behavior. According to the ERM, people's memory representations for categories consist of abstracted individual experiences with the category, called exemplars. When a test case is presented, a small sample of exemplars is retrieved, with the likelihood of retrieval being determined by how similar the stored exemplars are to the test case. The retrieved exemplars then enter an evidential assessment process in which the balance and magnitude of evidence are determined (Griffin & Tversky, 1992). Classification responses are governed by the balance of evidence, and likelihood judgments further depend on the magnitude of the balance. Sieck and Yates compared the ERM with a connectionist model of probability judgment, and found that the ERM provided superior accounts of overconfidence phenomena in three experiments. The ERM anticipates reductions in overconfidence under conditions that promote exemplar retrieval. In the common experimental task, participants learned to classify hypothetical patients with particular symptom patterns into disease categories and reported confidence judgments in the form of probabilities. One primary finding was that overconfidence decreased when a choice prompt was included, that is, when respondents were prompted to choose between diseases and report a 50–100% probability that the choice was correct, as compared with respondents who reported a 0–100% probability that each patient had the focal disease. A second principle finding was that overconfidence decreased with a direct instruction to retrieve many exemplars. The ERM, but not the connectionist model, predicted these phenomena.

In the current study, the ERM's proposal that choice inclusion effects on overconfidence are due to differences in amount of retrieval is tested against an alternative proposal that error in the response system is instead responsible (Juslin et al., 1997). A novel prediction of the ERM regarding the use of a relative frequency probe question is simultaneously tested. Specifically, the ERM predicts that reductions in overconfidence due to the use of a relative frequency probe, rather than a probability probe question, should be moderated by a choice inclusion manipulation. Finally, the model's ability to quantitatively fit likelihood judgment and choice data is examined for the first time. The plan of the remainder of this article is as follows. First, prior research on choice prompt inclusion and likelihood probe question effects are reviewed. Then, two formal process models are described, including their predictions regarding the interaction between choice inclusion and likelihood probe manipulations. An experimental test between the accounts is presented, and results and implications are then discussed.

CHOICE PROMPT INCLUSION

Confidence judgments are typically elicited by first requesting that the respondent choose the answer he or she believes is most likely to be correct, and then provide a 50-to-100% probability that the answer is in fact correct. Ronis and Yates (1987) compared judgment accuracy performance resulting from this assessment procedure with one in which respondents did not choose, but were instead asked to provide a probability,

ranging from 0 to 100%, that one of the two presented alternatives was correct. The former method was dubbed the Choice-50 (C50) procedure, and the latter was called the No-choice-100 (NC100) procedure. Ronis and Yates found greater overconfidence in general knowledge and basketball prediction tasks that employed the NC100 procedure than in the same tasks that used a C50 procedure.[1] These results were quite interesting, as they contrasted sharply with a proposal for overconfidence that was based on the self-perception theory for cognitive dissonance (Bem, 1967). A key tenet of self-perception theory is that free choice of a course of action increases the attractiveness of that action and decreases the attractiveness of other actions. This idea, as applied to overconfidence, clearly implies that more overconfidence should be found when a choice prompt is included. Sniezek et al. (1990) conducted a similar set of studies using general knowledge questions and found the same effects. Sniezek et al. explained their results by proposing that confidence becomes more appropriate as the amount of cognitive processing increases, and singled out information search processes in particular. Juslin and his colleagues have proposed an alternative to this explanation (Juslin et al., 1997, 1999). These researchers claim that interactions between the distinct scales used and random error in the response system are instead responsible, and have shown by simulation that such effects do follow from their proposal.

Following a similar line of thinking to that initiated by Sniezek et al., Sieck and Yates (2001) proposed that overconfidence in the C50 procedure is attenuated because the two prompts used in that procedure encourage people to engage in additional retrieval. They also showed by simulation that the additional retrieval is sufficient to produce the basic effect. In one experimental test of the retrieval idea, Sieck and Yates had respondents make judgments in a category learning task according to one of four conditions: (1) Standard/C50, (2) Standard/NC100, (3) Encode/NC100, (4) Recall/NC100. Recall/NC100 was the primary experimental condition, and participants therein were prompted with an instruction to retrieve as many of the cases they had seen on previous trials as possible prior to rendering their judgments using the NC100 procedure. Hence, respondents in this group produced only one judgment on the 100-point scale, as in the NC100 procedure, but they were prompted to retrieve additional information, as was claimed to occur in the C50 procedure. Participants in the Encode/NC100 condition followed a similar process, but were prompted with an instruction to examine each case carefully rather than to retrieve cases. The findings were that the Standard/C50 and Recall/NC100 conditions exhibited similar levels of overconfidence, and the Standard/NC100 and Encode/NC100 conditions were also similar to one another. However, less overconfidence was found in the Standard/C50 and Recall/NC100 conditions than in the Standard/NC100 and Encode/NC100 conditions. This result supports the hypothesis that retrieval is responsible for the effect, and the finding that the encoding instruction did not lead to a reduction in overconfidence implies that not any kind of additional processing is sufficient. The response error explanation also seems less likely in light of these data, though that explanation was not explicitly addressed. In the current study, the retrieval and response error hypotheses are directly compared.

---

[1]In the Ronis and Yates study, some respondents were given the opportunity to choose the most likely option and then to report the probability that their choice is correct on a 0–100% scale (a C100 procedure). These researchers found that participants sometimes reported probabilities less than 50% under this procedure, resulting in a reduction in overconfidence. Along the lines discussed by Ronis and Yates, this effect was perhaps due to the fact that the participants mistakenly believed that 0% represents maximal uncertainty, and the scale creates an experimental demand that fosters such a belief (In the participant's mind, 'Why would 0–50% be available, unless I am supposed to report them?'). Of course, 0% implies absolute certainty that the unchosen alternative is correct, whereas 50% represents maximal uncertainty. That is, the C100 procedure is problematic because it creates a conversational demand to report probabilities below 50%, which leads to choice and likelihood judgments that are incoherent. Evidence in support of this interpretation comes from Gigerenzer et al. (1991). They used a C100 procedure, but gave a very creative instruction about how to use it. Specifically, participants were told to use 0–50% whenever they wanted to change their original answer, rather than switching their choice selection, as they would presumably do otherwise. This kind of instruction eliminates the experimental demand by explicitly providing circumstances under which 0–50% ought to be used. And in contrast with the earlier research, Gigerenzer et al. found about a 0.01 (non-significant) increase in overconfidence in the C100 procedure, as compared with C50, as well as little use of confidence values below 50%. That is, with the demand eliminated, C100 produces results that are nearly identical to C50. Only C50 and NC100 procedures are considered in the current study, since both permit only coherent responses, and the models under investigation are not equipped to deal with issues as complicated as conversational demands.

LIKELIHOOD PROBE QUESTION

In addition to having participants report probabilities associated with each item, Sniezek et al. (1990) also collected estimates of the number of correct selections at the end of the general knowledge test. In contrast with the item-level confidence judgments, these frequency estimates exhibited underconfidence. Sniezek et al. suggested that differences in the types of evidence utilized in the item-level and aggregate judgments might explain the results, but noted that differences in the permissible scale ranges could also account for the effect. To elaborate on this latter possibility, consider that in a 2AFC task, participants who lack any knowledge of the material should expect to get half of the questions correct. However, such subjects might mistakenly believe that they will get all of the questions wrong. In such cases, nothing in the instructions or presented scale would dissuade these participants from reporting that they did not get any answers correct. More generally, consider respondents who believe that not knowing implies they are wrong, but nevertheless feel that they know many of the answers. Such respondents could produce judgments that underestimate their aggregate performance, and yet still exceed chance performance. That is, the phenomenon could exist at all points along the scale.

Gigerenzer et al. (1991) predicted reductions in aggregate frequency estimates as compared with item-level confidence judgments based on their Probabilistic Mental Models (PMM) theory. In PMM, if the answer to a question is not produced by memory and simple logic, then a probability structure representation that corresponds to a natural environment, called a reference class, is generated. The reference class determines what cues are available for a problem, and what their cue validities (conditional relative frequencies) are. Gigerenzer et al. proposed that item-level confidence judgments and aggregate frequency estimates would differ because distinct reference classes would be invoked in the two situations. For example, the reference class 'Cities in Germany' might be generated in response to an item-level question, whereas the class 'Sets of general-knowledge questions in similar testing situations' might be generated to aid in producing an aggregate frequency estimate. Since the latter class contains explicit experiences of incorrectly responding to tricky questions, confidence is lowered for that class. Furthermore, if the test questions are representative, then item-level confidence will be well calibrated and aggregate frequency estimates will exhibit underconfidence. If the questions are not representative such that tricky items are disproportionately sampled, then overconfidence will be found at the item level and good calibration will be found at the aggregate level.

In their first experiment, these researchers found essentially the same results as Sniezek et al., using quite similar procedures. And like Sniezek et al., they also noted that the effects could be due to the fact that item-level confidence judgments were limited to the upper half of the scale, whereas the aggregate frequency judgments were not. Gigerenzer and his colleagues attempted to deal with the problem in a second experiment by having some subjects report item-level confidence on a full-range scale. Unfortunately, they provided substantial instruction regarding the meaning and use of the item-level scale, but no corresponding instruction for the aggregate-level frequency scale. Hence, the scale interpretation remains a viable alternative explanation for the observed effects. More recently, Griffin and Buehler (1999) have found supporting evidence for this alternative explanation. Specifically, they had participants explain their aggregate frequency judgments, and then coded the protocols according to whether individuals explicitly accounted for the fact that they would get some answers correct merely by guessing. The critical results were that estimated frequency was substantially higher for those who explicitly adjusted for guessing as compared with those who did not, and in the former group, there was little difference between item-level confidence and estimated frequency. Thus, overconfidence differences that have been found between judgments regarding the number of items correct at the end of a general knowledge test, and probability judgments that each chosen alternative is correct seem to be primarily due to differences in the scales employed.

In the aforementioned studies, probability performance estimates regarding specific items were compared with frequency estimates of performance on a complete test. A couple of studies have compared probability and relative frequency probe questions at the item level, and also equalized the response scales (Brenner

et al., 1996; Price, 1998). For example, in a C50 general knowledge task, Price (1998) compared a standard probability probe question with the relative frequency question, 'Out of 100 questions for which you felt this certain of the answer, how many would you answer correctly?' According to PMM, a probe of this sort should lead to the generation of a 'test-question' reference class, and hence meaningful reductions in overconfidence ought to be found. Across three experiments, there was a consistent trend of about a 0.01 reduction in overconfidence for the groups that received the relative frequency probe (a meta-analysis that pooled the results of the three studies yielded a p-value of 0.07, implying that 0.01 is a firm estimate). Brenner et al. compared probability and relative frequency probes on a C50 personality prediction task and also found practically no difference between probe types on overconfidence.

Overall, the pattern of results speaks against the PMM proposal for differences between probability and relative frequency questions. The findings also imply that relative frequency probe question effects on item-level overconfidence will be minor or negligible, at least in general knowledge types of tasks. However, according to the ERM, a relative frequency question ought to produce meaningful effects in category learning tasks, at least in certain circumstances. Note that a relative frequency question, such as 'Out of 100 people like this one, how many would have the disease?' actually asks the participant to consider multiple cases prior to rendering a judgment. Hence, retrieval effects, such as were obtained by explicit instruction in earlier tests of the ERM, should result from a relative frequency question. However, as will be shown in the following section, the ERM predicts that probe question effects will be greater with NC100 than with C50 assessment procedures.

EXEMPLAR RETRIEVAL MODEL (ERM)

The ERM was described in detail by Sieck and Yates (2001), so it is summarized more briefly here. In the ERM, experiences with a category are stored separately in memory as exemplars. When a new case is presented, the respondent retrieves one or a few exemplars to inform their judgment. The probability that a particular exemplar is retrieved from memory is governed by the Medin and Schaffer (1978) retrieval rule. Imagine that the respondent has accumulated $j-1$ exemplars and is now facing case j (the probe). According to a single-parameter version of the Medin-Schaffer retrieval rule, the similarity between the probe, j, and the kth exemplar ($k = 1$ to $j-1$) is given by

$$\mathrm{sim}(j,k) = s^{d_k} \tag{1}$$

where d_k is the number of mismatching binary cues between the probe and exemplar, and s ($0 \leq s \leq 1$) is a parameter that represents the similarity of mismatching values for each symptom. Equation (1) says that the similarity between the exemplar and the probe decreases exponentially with each difference between them (Nosofsky, 1984; Shepard, 1987). The total probability that any exemplar from the target category, T, is retrieved is

$$\mathrm{P}(\text{retrieve T}\,|\,j) = \frac{\sum_{k\in T}\mathrm{sim}(j,k)}{\sum_{k\in T}\mathrm{sim}(j,k) + \sum_{k\notin T}\mathrm{sim}(j,k)} \tag{2}$$

$F_{T,N}$ denotes the respondent's personal probability that the probe belongs to the target category, based on the collection of exemplars retrieved. $F_{T,N}$ is the random variable given by

$$F_{T,N} = \frac{\alpha + \sum_{i=1}^{N} X_i}{\alpha + \beta + N} \tag{3}$$

where N represents the number of exemplars retrieved, and the X_i indicate each of the outcomes in the sample of retrieved cases such that $X_i = 1$ if the ith exemplar in the sample is a member of the target category, and $X_i = 0$ if the ith exemplar is not a member. Also, $\alpha/(\alpha + \beta)$ represents the respondent's personal probability of the target, prior to retrieving any past cases. This prior is not based on any real information, so $0 \leq \alpha = \beta \leq 1$, which form defensible Bayesian reference priors.

After the classification process has ended, the probe case presented on that trial is deposited into a long-term store. Two critical assumptions of the ERM are that each retrieval episode typically produces only one exemplar, but that additional prompting can induce retrieval of more exemplars. For example, in a C50 procedure, retrieval episodes are assumed to occur at both the prompt for a choice and the prompt for a confidence judgment (Sieck & Yates, 2001). An NC100 procedure entails only the prompt for a likelihood judgment, so a single retrieval episode occurs. This difference in the number of implicit demands to retrieve forms the basis for the ERM's account of reduced overconfidence in C50 as compared with NC100 procedures. More generally, rapid termination of retrieval provides the basis for the ERM's account of overconfidence. 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 set of retrieved exemplars points in the wrong direction. However, since 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.

The proposal that classification is often determined by the retrieval of a single exemplar can be traced back at least to Medin and Schaffer (1978), and has endured as a viable alternative in the categorization literature (Smith et al., 1998). Medin and Schaffer argued in favor of classification based on the first or first few exemplars retrieved, because classification based on an exhaustive retrieval plan would be implausible whenever large numbers of exemplars were stored in memory. Also, the multiplicative form of the retrieval equation implies that the retrieved exemplar will tend to be the probe's 'Nearest Neighbor', and classification on the basis of nearest neighbors often yields nearly optimal classification performance (Medin & Florian, 1992). Hence, reliance on a single case is fast and often effective. Furthermore, retrieval can be effortful, and quick termination reduces the costs associated with the effort expended (cf. Payne et al., 1992). An additional retrieval cycle at the confidence prompt is postulated because such a request constitutes a demand to produce novel information, which encourages people to reconsider the problem from a different perspective. The shift in perspective that results from the ensuing attempt to meet the demand alters the functional retrieval cues employed, so that more exemplars are retrieved. Research in memory retrieval has shown that large shifts in perspective can lead to additional recall via changes in the retrieval cues used (Anderson & Pichert, 1978). Further, research focused more specifically on repeated testing also suggests that added prompting leads to changes in the functional retrieval cues assembled, so that recall is increased (Roediger & Payne, 1982).

Overconfidence can be reduced by retrieval of more exemplars than is natural. However, it follows from the model that improvements in overconfidence due to retrieval will be met with diminishing returns. That is, the largest reduction in overconfidence should be found when two exemplars are retrieved, as opposed to one. Smaller reductions are expected with further additional retrieval. A simulation was conducted in order to illustrate these properties of the model. Figure 1 shows simulated overconfidence values for the model as a function of the number of exemplars retrieved. The simulation incorporated 2500 'participants' per data point and two values of s that constitute endpoints of the typical range for that parameter. The prior, α, was drawn from a uniform (0,1) distribution for each simulated participant in order to generate results based on a variety of underlying priors. In Figure 1, first note that overconfidence decreases as the number of retrieved exemplars increases. Observe also that the reductions in overconfidence diminish with increased retrieval. An implication of this feature of the model is that any two procedures that reduce overconfidence via increases in retrieval will produce effects that are subadditive.

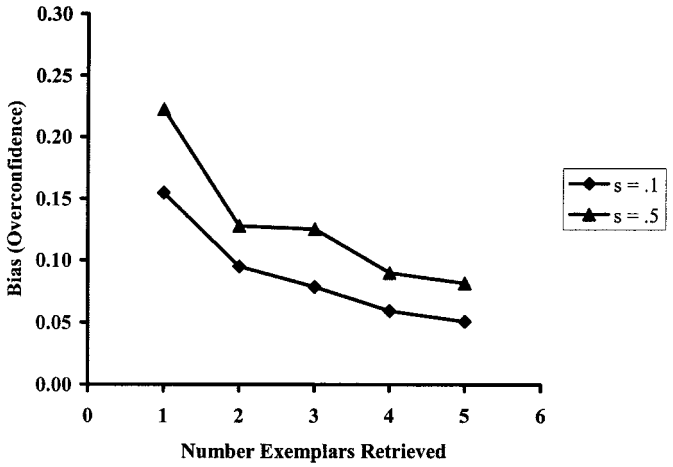

Figure 1. Simulated bias (overconfidence) from the exemplar retrieval model, as a function of number of exemplars retrieved and similarity (s)

EXEMPLAR RETRIEVAL AND RESPONSE ERROR MODEL (ERREM)

A crucial implication of the ERM described above is that, in a C50 procedure, confidence felt at choice differs from confidence felt at the time it is reported. This, in turn, accounts for the difference in overconfidence between C50 and NC100 assessment procedures. A natural and plausible alternative is that confidence is a simultaneous outcome of the choice process (e.g. Gigerenzer et al., 1991; Zakay, 1985). Under this assumption, the observed assessment procedure differences can be explained by interactions between the lengths of the respective scales and random variation in response processes (Juslin et al., 1997, 1999). A model with these distinct properties can be formulated from within the exemplar memory theoretical framework. And doing so is useful, as it allows for a comparison of the critical distinction between the models, while other model features are held constant (Estes, 1986).[2] Hence, the processing assumptions of the ERREM are quite similar to those of the ERM, and indeed equations (1)–(3) are preserved. The principle difference is that the ERREM proposes that reported confidence constitutes a noisy translation of internally experienced confidence. Reported confidence in one's choice is thus not a simple reflection of reported probability of the target about 50%. Specifically, in an NC100 procedure, $G_{T,N} = F_{T,N} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, denotes the respondent's reported probability that the probe belongs to the target category, based on the internal, personal probability, $F_{T,N}$. In a C50 procedure, the reported probability that a correct choice, C, was made is given by

$$G_{C,N} = \begin{cases} F_{T,N} + \varepsilon & F_{T,N} \geq 0.5 \\ 1 - F_{T,N} + \varepsilon & o.w. \end{cases} \tag{4}$$

Note that $G_{C,N}$ as resulting from equation (4) is not equivalent to the reflection of $G_{T,N}$ about 0.5, because the error term is added after the reflection in equation (4). This follows naturally from the definition of ε as noise

[2]This approach also produces a stronger competitor to the ERM than could be had by directly utilizing the combined-error model for confidence in general knowledge proposed by Juslin and his colleagues. For example, since the combined-error model was not developed in the context of category learning, it makes no claims about how people generalize from past observations to new cases. If it is assumed that there is no generalization, so that relative frequencies are encoded for each distinct pattern, then the model's performance on the task outstrips human performance. No model will do well in the current context without some mechanism for generalization.

arising in the motor response system. Also note that ε can push theoretical responses outside the permissible response range, in which case the corresponding endpoints of the range are substituted.

An important contribution of Juslin and his colleagues was to show that response error constructions of this sort are sufficient to obtain differences between C50 and NC100 assessment procedures. Hence, the ERM's assumption of post-choice retrieval may be unwarranted. In order to present a clear contrast between the models, the ERREM thus assumes that no additional retrieval occurs after choice, and that error in the response system operates as described above. In other respects, the model operates identically to the ERM.

A simulation was conducted in order to illustrate the C50 and NC100 predictions based on the response error assumption, and to reveal the interaction of such effects with exemplar retrieval. Figure 2 shows simulated overconfidence values for the model as a function of σ^2 and the number of exemplars retrieved. Other parameters were set as in the ERM simulation. Figure 2 illustrates that the response error component of the

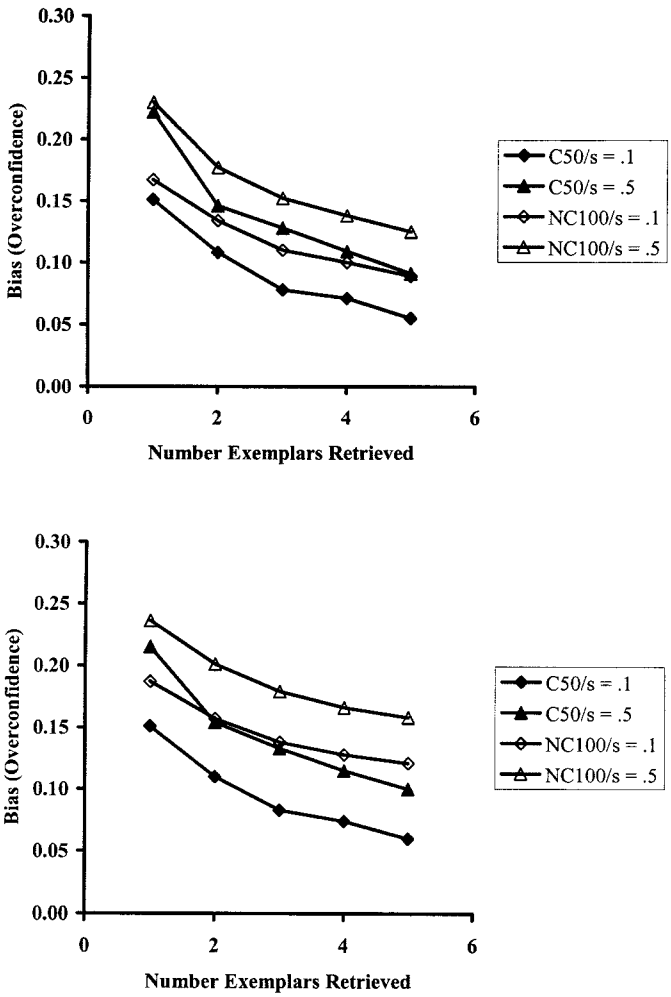

Figure 2. Simulated bias (overconfidence) from the exemplar retrieval and response error model, as a function of assessment task, number of exemplars retrieved, similarity (s), and response error (σ^2). Respondent chooses first and then reports 50–100% confidence that the choice is correct in the choice-50 (C50) assessment task. Respondent reports 0–100% likelihood that a specified event will occur in the no-choice-100 (NC100) task; the resulting likelihood values are converted to corresponding choice and confidence values. In the top panel, $\sigma^2 = 0.02$, and in the bottom panel, $\sigma^2 = 0.04$

ERREM allows the model to predict greater overconfidence with an NC100 than a C50 assessment procedure. The figure also shows that the NC100/C50 differences in overconfidence increase with exemplar retrieval. The effects due to response error are overshadowed by the effects due to retrieval when retrieval is highly abbreviated, but are brought to the fore as the retrieval problems are alleviated. This pattern of effects plainly contrasts with those of the ERM, thus enabling a sharp empirical test between the two models.

EXPERIMENT

The current experiment was conducted in order to examine the effects on overconfidence of choice prompt inclusion and likelihood probe question manipulations. The ERM and ERREM make opposing predictions about how these factors will interact with respect to overconfidence. According to the ERM, both factors impact overconfidence via differences in retrieval. The ERM anticipates diminishing returns in overconfidence improvements from increased retrieval. Diminishing returns would be indicated by an interaction, such that reductions in overconfidence due to inclusion of a choice prompt and incorporation of the relative frequency question are subadditive. Hence, according to the ERM, the choice inclusion effect should be larger for the probability probe question than for the relative frequency probe.

The ERREM, in contrast, proposes that distinct processes mediate the relations between each of these factors and overconfidence. Effects of choice inclusion are governed by interactions between random error in response processes and the different scale lengths employed. However, the possible numerical responses are identical for the probability and relative frequency probes. Thus, no likelihood probe differences are predicted on the basis of slips in confidence response. The ERREM anticipates that the choice inclusion effect should be at least as large with a relative frequency probe as with a probability probe. To the extent that a relative frequency probe promotes retrieval, the choice inclusion effect should be larger for that group, as the subtler response error effects become more apparent once retrieval limitations have been alleviated. That is, the two effects should be superadditive. The ERM and ERREM make divergent predictions regarding the nature of the interaction, thus allowing a powerful test between the models. The contrasting predictions of these two models are evaluated in the following experiment.

METHOD

Participants

Study participants were 121 undergraduate students enrolled in an introductory psychology course at the University of Michigan. Experimental participation was part of their course requirement.

Stimuli

The stimuli took the form of text descriptions of individual 'patients'. Each patient was identified by a unique set of two letters that represented the patient's initials, and was described as having or not having the symptoms, rash (R), cramps (C), and red eye (E). Each of the 140 patients belonged to one of two hypothetical disease categories: 'Trebitis' or 'Philiosis.' The base rates for Trebitis and Philiosis were 0.50/0.50. The symptoms were arranged into seven uniformly distributed symptom patterns that included all possible combinations of individual symptoms, except for the no-symptom possibility. The conditional relative frequencies of Trebitis for each pattern are displayed in the second column of Table 1.

Design

The experiment was a 2 (choice prompt inclusion) $\times$ 2 (likelihood probe question) between-subjects design. The two crossed factors provided four methods of elicitation of the probabilistic differential diagnosis. One

Table 1. Predicted and observed choice proportions for each elicitation condition

		C-P50		C-R50	
Symptom pattern	P(T)	Pre.	Obs.	Pre.	Obs.
E	0.10	0.309	0.256	0.227	0.221
C, E	0.20	0.364	0.350	0.301	0.221
R, E	0.40	0.472	0.413	0.457	0.417
C	0.50	0.537	0.541	0.556	0.448
R, C, E	0.60	0.530	0.622	0.544	0.710
R	0.80	0.672	0.728	0.747	0.793
R, C	0.90	0.701	0.666	0.785	0.714
		NC-P100		NC-R100	
Symptom pattern	P(T)	Pre.	Obs.	Pre.	Obs.
E	0.10	0.309	0.200	0.227	0.171
C, E	0.20	0.364	0.369	0.301	0.219
R, E	0.40	0.472	0.469	0.458	0.471
C	0.50	0.537	0.434	0.556	0.435
R, C, E	0.60	0.530	0.645	0.544	0.729
R	0.80	0.672	0.600	0.747	0.652
R, C	0.90	0.701	0.679	0.785	0.629

Note: P(T) = relative frequency of the target disease ('Trebitis'), given the pattern. The individual symptom abbreviations are R = rash, C = cramps, and E = red eye. C-P50 = Choice-Probability-50, C-R50 = Choice-Relative Frequency-50, NC-P100 = No-Choice-Probability-100, NC-R100 = No-Choice-Relative Frequency-100. Pre. = predicted by ERM; Obs. = observed.

was the standard 'Choice-Probability-50' (C-P50) 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 respond to the question: 'What's the probability that this patient has the disease you indicated?' Participants following the second, 'Choice-Relative Frequency-50' (C-R50), method also indicated the most likely disease, and then responded to the question: 'Out of 100 people like this one, how many would have the indicated disease?' The third method was the 'No-Choice-Probability-100' (NC-P100) assessment procedure, in which participants responded to this question for each patient: 'What's the probability that the patient has Trebitis (rather than Philiosis)?' The fourth method was the 'No-Choice-Relative-Frequency-100' (NC-R100) elicitation procedure, wherein participants responded to the question: 'Out of 100 people like this one, how many would have Trebitis (rather than Philiosis)?'

Procedure

The experiment was conducted entirely via computer. The study participants were instructed to imagine that they were physicians faced with the problem of learning to diagnose between two new diseases that have appeared in their community. They were told that the patients would either have or not have each of the three symptoms, that they would have to learn the relationships between symptoms and diseases, that the symptoms were only imperfect predictors of disease category, and that they would be asked to provide likelihood information as a part of their diagnosis procedure. Participants were also given specific instructions concerning use of the likelihood scale. Participants in the C-R50 condition were instructed to adhere to the following conventions when stating their likelihood judgments:

(a) 50 should mean that half of the patients with similar symptoms would have Trebitis, and half would have Philiosis.

(b) 100 should mean that all of the patients with similar symptoms would have the disease you indicated for your first judgment.
(c) Increasing frequencies between 50 and 100 should correspond to increasing numbers of similar patients actually having the medical condition you stated.

Participants in the NC-R100 condition were told to adhere to these conventions:

(a) 50 should mean that half of the patients with similar symptoms would have Trebitis, and half would have Philiosis.
(b) 100 should mean that all of the patients with similar symptoms would have Trebitis, and 0 should mean that all of the patients with similar symptoms would have Philiosis.
(c) Increasing frequencies between 50 and 100 should correspond to increasing numbers of similar patients actually having Trebitis rather than Philiosis.
(d) Decreasing frequencies between 50 and 0 should correspond to increasing numbers of similar patients actually having Philiosis rather than Trebitis.

Participants in the C-P50 and NC-P100 conditions were instructed to adhere to corresponding conventions that were phrased in terms of probabilities for the given patient, rather than frequencies of similar patients. They were identical to the conventions stated in Sieck and Yates (2001).

On each trial, the participant: (1) was presented with a new patient, who was identified by two initials, and that patient's symptom profile; (2) indicated a probabilistic differential diagnosis according to one of the four elicitation procedures, as described above; and (3) received feedback about what was 'eventually determined' to be the patient's actual condition to permit learning. Subjects made diagnoses for 70 patients during an initial block of trials and then diagnosed another 70 patients in a second block after a short rest break. Analyses focus on Block 2, after some learning has occurred, because those judgments are of primary importance for the hypotheses under consideration.

RESULTS

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 non-focal 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 non-focal alternative was selected.

Empirical observations

The empirical data are described in this section, and theoretical explanations are provided in the subsequent section.

Choice proportions and mean likelihood judgments

The observed choice proportions for each symptom pattern are shown in Table 1, and corresponding mean likelihood judgments are shown in Table 2. Both the observed choice proportions and mean judgments tended to overestimate the actual relative frequencies of the target when those relative frequencies were below 0.5, and underestimate the actual relative frequencies when they were greater than 0.5. Note that this effect is more pronounced for the likelihood judgments than for the choice proportions.

Table 2. Predicted and observed mean likelihood judgments for each elicitation condition

		C-P50		C-R50	
Symptom pattern	P(T)	Pre.	Obs.	Pre.	Obs.
E	0.10	0.360	0.343	0.340	0.363
C, E	0.20	0.401	0.416	0.387	0.404
R, E	0.40	0.479	0.484	0.476	0.481
C	0.50	0.527	0.485	0.531	0.479
R, C, E	0.60	0.521	0.536	0.525	0.579
R	0.80	0.625	0.587	0.644	0.637
R, C	0.90	0.648	0.559	0.669	0.565
		NC-P100		NC-R100	
Symptom pattern	P(T)	Pre.	Obs.	Pre.	Obs.
E	0.10	0.394	0.267	0.349	0.294
C, E	0.20	0.424	0.388	0.392	0.364
R, E	0.40	0.484	0.501	0.478	0.451
C	0.50	0.521	0.427	0.530	0.490
R, C, E	0.60	0.516	0.603	0.523	0.643
R	0.80	0.595	0.607	0.635	0.606
R, C	0.90	0.612	0.621	0.659	0.580

Note: See Table 1.

Proportion correct

The observed proportions of correct diagnoses are shown in the upper panel of Figure 3 for each condition. As can be seen, the relative-frequency groups achieved slightly higher proportions correct than the probability groups, $F(1, 117) = 5.41$, $p = 0.022$. But choice prompt inclusion and the interaction were not significant.

Bias (over-/underconfidence)

Over-/underconfidence was indexed, as is the norm, via the following bias statistic:

$$\text{Bias} = \text{mean probability judgment} - \text{proportion correct} \tag{4}$$

Positive values for bias indicate overconfidence, and negative values indicate underconfidence. The upper panel of Figure 4 shows the mean values of the bias statistic for each condition. The main effects for likelihood probe and choice inclusion, as well as the interaction, were statistically significant, $F(1, 117) = 8.17$, $p = 0.005$, $F(1, 117) = 11.77$, $p = 0.0008$, and $F(1, 117) = 5.34$, $p = 0.023$, respectively. As illustrated, the choice inclusion effect for the probability probe question is substantially larger than that for the relative frequency probe, thus revealing subadditivity for the combined effects. As can also be seen, overconfidence in the C-R50 condition is still significantly positive, $t(28) = 3.54$, $p = 0.001$, implying that the combination of the two procedures does not readily eliminate overconfidence, so that there is still considerable room for improvement. This additionally indicates that the interaction is not due to a floor effect (the minimum possible value of bias for the proportion correct achieved in C-R50 is -0.09, or 0.17 points lower than the actual value). Finally, a significant difference was found for proportion correct, so one may consider the hard–easy effect (i.e. that less overconfidence is often associated with a higher proportion correct) as an explanation for the overconfidence findings. However, the main effect found for proportion correct cannot explain the

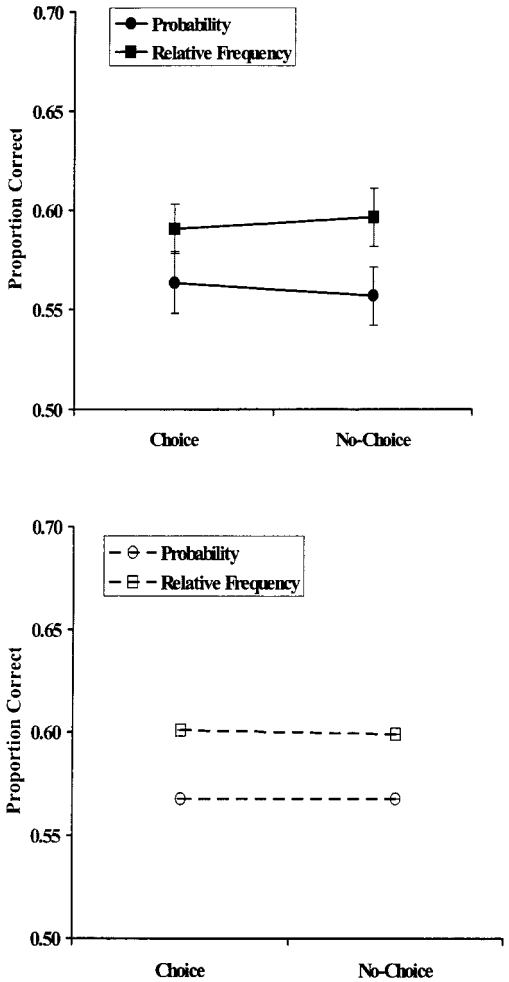

Figure 3. Proportion correct for each elicitation condition: Observed and Predicted by ERM. Empirical observations are shown in the top panel, and theoretical predictions are shown in the bottom panel

interaction found for overconfidence. As will be seen in the next section, the mechanisms of the ERM are sufficient to quantitatively account for the complete pattern of effects.

This pattern of results is consistent with the ERM, but contrasts sharply with the ERREM's predictions. Specifically, the ERREM's assumption that choice inclusion effects result from variability arising in the response system cannot account for the observed interaction effect. Since the ERREM incorrectly predicts the ordering of effects, it will necessarily yield worse quantitative predictions than the ERM. However, the ERM will not necessarily produce reasonable quantitative fits to the data, just because it correctly predicted the general pattern of results. The ERM's ability to fit the data quantitatively is thus assessed in the following section.

Quantitative predictions of the ERM

The ERM was fitted simultaneously to the choice proportions and mean likelihood judgments observed for each distinct symptom pattern. That is, the model was fitted to the data in Tables 1 and 2. In fitting

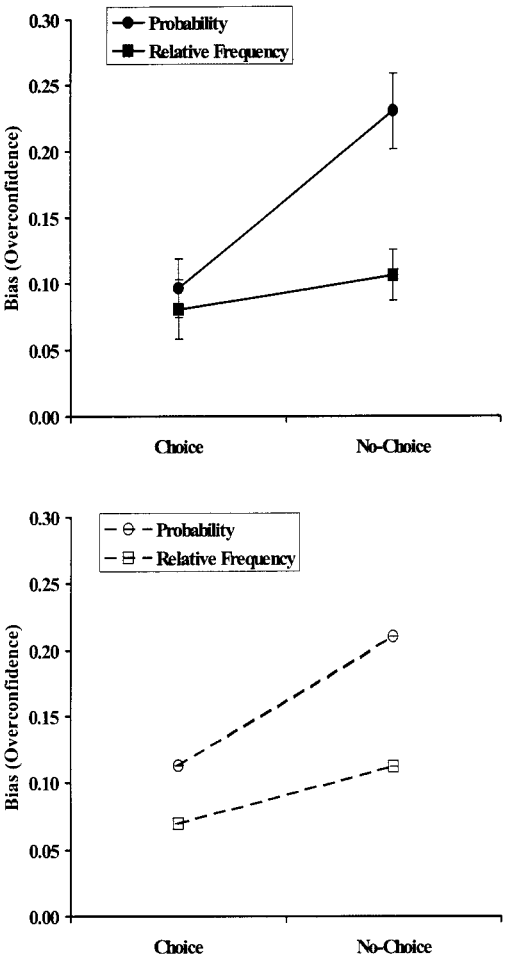

Figure 4. Overconfidence for each elicitation condition: Observed and Predicted by ERM. Empirical observations are shown in the top panel, and theoretical predictions are shown in the bottom panel

the model, it was assumed that an additional exemplar was stored in memory upon presentation of the feedback at the end of each trial. Also, the model was first used to generate choice probabilities and mean likelihood judgments for each symptom pattern on each individual trial. These individual trial predictions were then averaged to create the grouped symptom pattern predictions that were compared with the observed data.

The data from the four conditions were simultaneously fit by six free parameters: a single value of the similarity parameter, s, and a single value of α (which forms the prior) were used, along with four retrieval parameters, N, that varied by condition. The ERM was fitted to the data by conducting a computer search for the free parameters that minimized the root mean squared error (RMSE) between predicted and observed choice proportions and mean likelihood judgments. The best-fitting parameters and statistics on the fits are reported in Table 3. To my knowledge, explicit fits to both choice proportions and mean likelihood judgments have not been previously reported. Nevertheless, as shown in Table 3, the predicted values correlate fairly well with observations. Also, the fit could be improved by lessening the restrictions on the parameters, for example, by allowing distinct values of s for each symptom.

Table 3. Statistics on fits of ERM to choice proportions and mean likelihood judgments

		Choice proportion		Mean likelihood	
Condition	*N*	RMSE	*r*	RMSE	*r*
C-P50	3	0.0527	0.954	0.0411	0.947
C-R50	5	0.0882	0.920	0.0498	0.906
NC-P100	1	0.0769	0.890	0.0699	0.887
NC-R100	3	0.1147	0.856	0.0627	0.868

Note: $s = 0.378$ and $\alpha = 0.400$ for all conditions. RMSE = root mean squared error, r = correlation coefficient.

Choice proportions and mean likelihood judgments

Predictions of the choice proportions and mean likelihood judgments are shown along with the observed data in Tables 1 and 2. As noted above, the predictions conform reasonably well to the actual observations. The model expects the finding that choice proportions more closely resemble probability matching than probability maximizing because choices are based on small samples of past exemplars, and because retrieved exemplars will not necessarily identically match the probe case. For example, some patients with Cramps and Red Eye might be retrieved and used when the probe patient has only Red Eye.

As noted above, likelihood judgments are drawn more closely towards 50% than choices. The model expects this effect because, first, as with choices, the past exemplars from other covariance classes might be retrieved. And there is an additional effect of the prior, α, which is to pull likelihood judgments in towards 50%. That is, rendered judgments are systematically less extreme than proportions from retrieved samples of cases, depending on a global degree of ambivalence about the task which is felt before any case specific information is retrieved. Finally, at a finer grain, allowing for more free parameters would improve the fit. For example, the model would better capture some of the non-monotonicities evident in the data if independent *s* values were used for each symptom.

Proportion correct

The predicted proportions of correct diagnoses were derived by first using the best fitting parameters to generate individual trial predictions, and then averaging over trials. They are shown in the lower panel of Figure 3 for each condition. As can be seen, the model mimics the finding that the relative-frequency groups achieved higher proportions correct than the probability groups. Clearly, the model predicts a higher proportion correct with greater retrieval. It is especially interesting that the model captured the difference between the NC-R100 and C-P50 groups, even though three exemplars were estimated to have been retrieved for each of these groups (and other parameters were held constant). The reason is that all three exemplars would be retrieved prior to judgment, and so factored into the choice, in the NC-R100 task. In the C-P50 task, the third exemplar would be retrieved at the confidence prompt, and so would not influence choice.

Bias (over-/underconfidence)

The lower panel of Figure 4 shows the predicted mean values of the bias statistic for each condition. Again, the pattern of effects is well captured by the model. That is, a positive bias is found in all conditions, and the choice inclusion effect is larger for the probability probe than for the relative frequency probe question. Recall that the model predicts overconfidence in general because of retrieval limitations. It also predicts the interaction because improvements in overconfidence exhibit diminished returns as the number of retrieved exemplars increases.

In summary, in addition to accounting fairly well for the choice proportions and mean likelihood judgments, the ERM correctly predicted the main effect for proportion correct, as well as the interaction found for overconfidence.

DISCUSSION

Two distinct exemplar memory models were compared on their ability to predict the individual and joint impact of two factors on overconfidence in the context of category learning: (1) choice prompt inclusion, and (2) likelihood probe question. The relevant findings and implications regarding each of these factors are discussed in turn.

Choice prompt inclusion

Overconfidence is reduced when respondents are asked to first choose, and then report the probability that their choice is correct, rather than being asked to report the probability that a pre-specified alternative is correct. Since its initial establishment by Ronis and Yates (1987), this effect has proved to be quite general and reliable. Consequently, any model of likelihood judgment demanding serious attention ought to offer some explanation for this choice inclusion effect. Two distinct proposals for this effect that have recently emerged in the literature were tested in the current study. Specifically, the ERM's proposal that the choice inclusion effect on overconfidence is due to differences in amount of retrieval was tested against the assumption of an alternative model (ERREM) that such differences are due to random error in the response system. Experimental results clearly indicated that the choice inclusion effects were substantially larger when a probability probe question was used than when a relative frequency probe was employed. This result is consistent with the underlying mechanisms of the ERM, but not with those of the ERREM. Specifically, the ERREM's proposal that choice inclusion effects result from response error cannot account for the observed interaction effect. The numeric scales were identical in the two likelihood probe conditions. Consequently, there is no reason on the basis of response error to expect the choice inclusion effect to be reduced when a relative frequency probe is adopted. More generally, the choice prompt difference for the relative frequency question can be taken as an upper bound on the size of effects to be expected from response error and other scaling considerations. These findings propose a serious challenge to theories that assume choice and confidence are determined simultaneously, since such theories are seemingly limited to scaling proposals to account for choice inclusion effects.

Likelihood probe question

Frequency and relative frequency probe questions have received considerable attention in the literature in the past few years, though most of the work has been centered on post-test estimates of aggregate performance. A couple of studies have compared effects of probability and relative frequency probe questions on overconfidence at the item level using a C50 procedure, and the results of these studies indicate essentially nil effects of likelihood probe (Brenner et al., 1996; Price, 1998). Yet, the ERM anticipated that in an NC100 procedure a relative frequency probe question would reduce overconfidence, as such a question literally demands the consideration of multiple cases. Choice inclusion was expected to moderate the relative frequency effect, since both factors are governed by retrieval according to the ERM, and the benefits of retrieval exhibit decreasing returns. The results of the current experiment bore these predictions out. In addition to correctly anticipating these novel findings qualitatively, the ERM also provided good quantitative fits to the data.

The finding that a relative frequency probe reduces overconfidence when no choice prompt is included is theoretically interesting. The result provides encouraging support for the underlying mechanisms of the ERM, and presents a new challenge for extant and future models. Nevertheless, from a practical perspective, relative frequency elicitation did not offer any real improvement over and above choice inclusion. As such, the current research is consistent with other recent studies indicating that relative frequency questions will not provide easy solutions to our judgment problems (cf. Griffin & Buehler, 1999). If relative frequency probes do not readily produce ideal judgment, what steps can be taken to yield practically significant improvements? This issue is discussed presently from the perspective of the ERM.

Implications for probability judgment expertise

Keren (1987) proposed that ample practice in judgment tasks with highly related events and regular feedback should lead to good calibration. The knowledge representation and retrieval assumptions of the ERM imply that, although they may be necessary, these criteria will generally be insufficient. Specifically, in high-variance domains, the information that can be derived from one or a few retrieved cases is inadequate to support the formulation of reliable judgments. This would be true even for an expert who might have thousands of exemplars stored in memory. Given these retrieval limitations, how could an expert develop excellent calibration? One way would be to routinely utilize (first inventing, if necessary) high-quality analytical procedures for 'scoring' cases according to their strength along the criterion. As experience with such a procedure increases, more and more episodes that contain both case information and the score values are encoded. And retrieval of a small number of these 'enriched' exemplars would yield information that could adequately support likelihood judgment, even without the analytical procedure. This conceptualization may help explain why the calibration of some experts is exceptional, while it is quite poor for others.

First, consider probability forecasting in meteorology, where outstanding calibration is the norm. Since the early 1900s, when meteorologists began experimenting with the formulation of probabilistic forecasts, they have also worked to develop (initially simple) objective procedures to aid their judgment (Murphy & Winkler, 1984). Naturally, objective categorical and probabilistic forecasts remain indispensable inputs for subjective probability forecasts in meteorology. Extremely good calibration has also been found among experts in bridge (Keren, 1987). Bridge consists of two phases, bidding and play. Keren had participants assess the likelihood that the final contract reached during bidding would be made during play. An essential aspect of the bidding phase is the assessment of the strength of one's hand, and this is not something that is left to unaided intuition. Instead, analytical procedures that have been studied and practiced are used to score the strength of the hand, and other systematic routines are employed to best communicate hand strength to one's partner. Hence, analytical guides are regularly incorporated into the judgment process in both of these spheres. This incorporation clearly depends, in turn, on two necessary conditions. First, such routines must be available, and second, judges must be willing to adopt them.

Now turn to a field where calibration is quite poor, such as medicine (Christensen-Szalanski & Bushyhead, 1981; Dawson et al., 1993). In this case, the analytical routines may be available, but they are not willingly adopted (e.g. Graham et al., 2001). Substantial research in diagnostic judgment in clinical psychology points to the same general conclusion. Professionals in both of these fields prefer to rely on unaided 'clinical' experience when making diagnostic judgments, rather than formal procedures. And their diagnostic judgment performance suffers for it (Dawes et al., 1989; Goldberg, 1968). Again, the ERM implies that, even with regular outcome feedback, experience alone will be insufficient to achieve competence in calibration because of retrieval limitations. On the positive side, the incorporation of analytical procedures in addition to (or even in lieu of) regular outcome feedback is not only expected to aid immediate judgment, but also to aid in the development of a stronger knowledge base. Extensive studies in both simulated and real judgment situations are needed to test these claims.

ACKNOWLEDGEMENTS

This work was supported in part by Grants SES-9911301 and SBR-0196200 from the National Science Foundation. The author would like to thank Hal Arkes, Edgar Merkle, and Paul Price for comments on previous versions of this paper. This work also benefited from many excellent discussions about confidence with Frank Yates.

REFERENCES

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17*, 1–12.

Bem, D. J. (1967). Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74*, 183–200.

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.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: how can experts know so much and predict so badly? In K. A. Ericsson, & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge: Cambridge University Press.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilisitic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 928–935.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Dawson, N. V., Connors, A. F., Jr., Speroff, T., Kemka, A., Shaw, P., & Arkes, H. R. (1993). Hemodynamic assessment in managing the critically ill: is physician confidence warranted? *Medical Decision Making, 13*, 258–266.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130*, 579–599.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18*, 500–549.

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

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*, 483–496.

Graham, I. D., Stiell, I. G., Laupacis, A., McAuley, L., Howell, M., Clancy, M., Durieux, P., Simon, N., Emparanza, J. I., Aginaga, J. R., O'Connor, A., & Wells, G. (2001). Awareness and use of the Ottawa Ankle and Knee Rules in 5 countries: can publication alone be enough to change practice? *Annals of Emergency Medicine, 37*(3), 259–266.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: easy solutions to cognitive illusions? *Cognitive Psychology, 38*, 48–78.

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

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*, 279–285.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1–15.

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

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(24), 771–792.

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

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

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79*(387), 489–500.

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

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

Price, P. C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: the case of external correspondence. *Organizational Behavior and Human Decision Processes, 76*(3), 277–297.
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.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: a comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*(4), 1003–1021.
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.
Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*(2), 89–117.
Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica, 58*, 75–80.

Author's biography:

Winston Sieck received his PhD in cognitive psychology from the University of Michigan in 2000, and is currently a post-doctoral scholar in quantitative psychology at Ohio State University. His research interests include confidence, memory, and argumentation.

Author's address:

Winston R. Sieck, Department of Psychology, Ohio State University, 1827 Neil Avenue, Columbus, OH 43210-1222, USA.

