INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

INTEGRATION OF THE ECOLOGICAL AND ERROR MODELS OF OVERCONFIDENCE USING A MULTIPLE-TRACE MEMORY MODEL

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

MICHAEL R. P. DOUGHERTY

Norman, Oklahoma

1999

INTEGRATION OF THE ECOLOGICAL AND ERROR MODELS OF OVERCONFIDENCE USING A MULTIPLE-TRACE MEMORY MODEL

A Dissertation APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

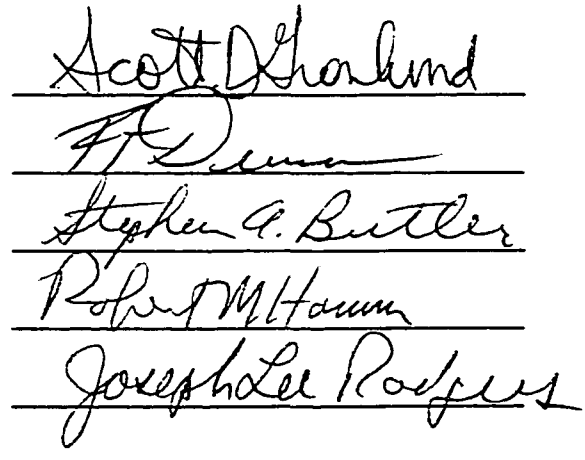


Table of Contents

Abstract

This research examined a memory processes account of the calibration of probability judgments. A multiple-trace memory model, MINERVA-DM (MDM; Dougherty, Ogden, & Gettys, 1999), was used to integrate the ecological (Brunswikian) and the error (Thurstonian) models of overconfidence. The model predicts that overconfidence should decrease both as a function of experience and as a function of encoding quality. Both increased experience and improved encoding quality result in lower error variance in the output of the model, which in turn leads to better calibration. Three experiments confirmed these predictions. Implications of MDM's account of overconfidence are discussed.

Considerable research has been devoted to understanding the accuracy (or inaccuracy) of probability judgments. Most recently, this research has been dominated by the calibration paradigm, in which the accuracy of people's probability judgments is assessed by comparing their judgments to their proportion of correct inferences. In the typical experiment, participants answer a series of general knowledge questions such as "What is the capital of Brazil? A) Sao Paulo B) Brazilia," and then state their confidence in their answer. Accuracy is assessed by measuring the calibration of participants' judgments. A person is said to be well calibrated if the average of their confidence judgments equals their proportion of correct judgments (confidence = proportion correct). Miscalibration, on the other hand, is characterized as either overconfidence (confidence > proportion correct), or underconfidence (confidence < proportion correct). The typical finding in the calibration paradigm is that people are overconfident, giving probability judgments that exceed their proportion of correct inferences (Lichtenstein, Fischhoff, & Phillips, 1982).

Despite the considerable amount of research on the overconfidence effect, the locus of overconfidence is still disputed. One theoretical account of overconfidence, the Brunswikian account, suggests that it is an artifact of the experimental task. Proponents of the Brunswikian account (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991; Juslin, 1994; Juslin, Olssen, & Winman, 1997) argue that studies investigating the calibration of probability judgments must meet two related pre-conditions: 1) the to-be-judged questions need to be drawn from an ecological reference class familiar to the participant, and 2) the sample of questions used should be drawn randomly from the

reference class. The ecological reference class is a function of the environment in which people interact. To the extent that people have experience with an ecological reference class, their mental representation of that reference class will be more-or-less veridical. Therefore, if questions are drawn from a person's ecological reference class, and the assumption of random sampling is met, participants should give judgments that are more-or-less well calibrated. Gigerenzer et al. argued that most studies investigating overconfidence have not met these two conditions. In fact, several studies now show that overconfidence often is reduced considerably when the ecological reference class and random sampling assumptions are met (see Juslin, Olsson, & Björkman, 1997 for a review; but see Brenner, Koehler, Liberman, & Tversky, 1996; and Griffin & Tversky, 1992; for counterexamples).

A second major theoretical account of overconfidence is the Thurstonian, or random error account (Erev, Wallsten, & Budescu, 1994), which assumes that overconfidence results from internal cognitive processes. Specifically, the Thurstonian account suggests that overconfidence is the result of random error associated with the response process. Using a model based on signal detection theory, Erev et al. showed that overconfidence could be accounted for by assuming that true judgments were perturbed by random error. Erev et al.'s model predicted that overconfidence should increase as the random error associated with the response process increased. Thus, in contrast to the Brunswikian account, the Thurstonian account attributes the overconfidence effect to random noise in the cognitive system.

Whereas both the Brunswikian and Thurstonian approaches to overconfidence have their merits, neither approach specifies the cognitive processes that give rise to both good and poor calibration. The Brunswikian account posits that the locus of the overconfidence effect is the result of factors external to the decision maker: Overconfidence is assumed to arise from the structure of the environment and the biased selection of questions on the part of the experimenter. The Thurstonian account, in contrast, suggests that the locus of overconfidence is the result of internal cognitive processes (i.e., response processes), but it fails to identify the precise nature of the processes that lead to the random error (however see Wallsten, Bender, & Li, 1999; Wallsten & Gonzalez-Vallejo, 1994 for a discussion of some of these issues). Recent research investigating these models suggests that neither account is sufficient to account for overconfidence (Budescu, Wallsten, & Erev, 1997; Juslin et al., 1997), and that a combination of the Brunswikian and Thurstonian models may be necessary.

The present research examines an alternative to the above theoretical accounts of overconfidence that is based on a mathematical memory model, MINERVA-DM (MDM; Dougherty, Gettys, & Ogden, 1999). MDM provides a memory processes model for studying overconfidence that capitalizes on both the Thurstonian idea of random variation, and the Brunswikian idea of ecological structure (See Björkman, 1994; Juslin, Olssen, & Björkman, 1997; and Soll, 1996 for models similar in spirit). However, unlike the Thurstonian and Brunswikian models, MDM makes explicit the memory processes and the representational assumptions that underlie judgments of probability and confidence, and it makes new predictions regarding the factors that lead

to both good and poor calibration. MDM is not offered as a competing model of overconfidence, but rather as an integrative model that captures the spirit of both the Thurstonian and Brunswikian approaches. The next section briefly describes MDM and how it incorporates both the Thurstonian and the Brunswikian assumptions. The description of MDM is brief; readers interested in a more thorough treatment of the model are referred to Dougherty et al. (1999).

Integrating the Brunswikian and Thurstonian Approaches Using MDM

Overview of MDM.

MDM is a modified version of Hintzman's (1988) MINERVA2 memory model, and as such, retains all of MINERVA2's original capabilities as a memory model. The primary modification to MINERVA2 is the incorporation of a two-stage conditional memory retrieval process that enables the model to account for a wide array of judgmental phenomena, including conditional probability and frequency judgments, and several of the heuristics and biases (see Dougherty et al., 1999).

Two core properties of MDM are its assumptions about how information is encoded and represented in memory and its assumptions about retrieval. MDM is an instance-based model (Dougherty et al., 1999; see also Hintzman, 1988) and it is assumed that a new memory trace is encoded in memory for each experienced event. Encoding in MDM proceeds by creating long-term memory traces by copying an event vector. The event vector is a representation of the external, environmental event (for mathematical purposes both the event vector and the memory traces are represented as vectors of +1's, 0's and -1's). An important assumption of MDM is that traces are

degraded versions of the event vector that created them. This degradation process is modeled by the encoding parameter L (which varies from 0 to 1.0). With higher values of L, better copies of the event vector are stored in memory; with lower values of L, worse copies of the event vector are stored in memory. The memory representation in MDM is assumed to consist of a database of instances representing the decision maker's past experiences and these instances are assumed to contain the components necessary to model a variety of judgmental phenomena. For generality, it is assumed that instances contain up to three concatenated components: a hypothesis component (H), a data component (D), and an environmental context component (E). The H component represents any event about which we wish to make a judgment, the D component represents the information or data that are used as input for that judgment, and the E component can be used to represent either, or both, intra and extra-experimental context. Importantly, the H and D components form the basis of the MDM inference engine, as they enable the model to account for conditional probability judgments, such as P(H|D) or P(D|H), and non conditional, or base-rate, judgments such as P(H) and P(D).[1]

Retrieval proceeds by computing the similarity between a probe vector and each trace stored in memory using

$$S_i = \frac{\sum_{j=1}^{N} P_j T_{i,j}}{N_i} \qquad (1)$$

[1] Although the environmental context component has been useful for simulating the effect of extra-experimental memory traces on judgments of frequency, it has yet to be fully investigated in the context of conditional judgments. It will not be used in any of the simulations presented in this article.

where $\mathbf{P}_j$ represents feature $_j$ in the probe vector, $\mathbf{T}_{ij}$ represents feature$_j$ in trace vector i, and N_i is the number of corresponding nonzero features in both the probe and trace. The activation for each trace, A_i, is given by cubing its similarity,

$$A_i = S_i^3. \tag{2}$$

These cubed similarities are then summed over all traces stored in memory to give the overall echo intensity, I,

$$I = \sum_{i=1}^{M} A_i \tag{3}$$

where M is the total number of instances stored in memory. It is assumed that I is proportional to the judged frequency or probability for non-conditional judgments such as P(H): $I \approx$ P(H).

The retrieval process for conditional probability judgments is slightly different, and is assumed to proceed in two stages. Imagine you are asked to judge P(H|D), for example, the probability that someone is a democrat (H) given that she is wealthy (D), P(democrat|wealthy). In this judgment, the decision maker is interested in estimating the frequency of H's in the subset of D's stored in memory. Thus, the first stage of this process entails delineating the subset of instances stored in memory corresponding to the condition, D. This is done by computing the similarity between the D portion of the probe and the D portion of each trace stored in memory to determine which traces correspond to the particular type wanted. For example, if the condition is "wealthy," then the similarity between the "wealthy" component of each trace in memory and the wealthy component of the probe is computed. If the similarity of the trace meets or exceeds the threshold criterion, S_c, it is assumed to be passed on to the second part of

the retrieval process. However, if the similarity value is less than S_c, processing of that trace stops. S_c is a threshold parameter that determines the minimum amount of similarity needed for a particular trace to be placed in the activated subset. Thus, the first stage of the process delineates, or activates, a subset of instances in memory on which the conditional probability is based.

The second stage of the process entails the assessment of the relative frequency of traces in the delineated subset that correspond to the non-conditional part of the judgment. For example, in the democrats given wealthy judgment, how many traces in the activated subset of wealthy people correspond to "democrats?" This is achieved by computing the similarity between the H component of the probe and the H component of each trace in the activated subset, cubing the similarity value for each trace, and summing the cubed similarities across all traces in the activated subset. This gives rise to the *conditional echo intensity*, I_c, which is given by

$$I_C = \frac{I_{S_i \geq S_C}}{K} \tag{4}$$

where K is a count of the number of traces that passed the S_c criterion value. For conditional probability judgments, it is assumed that I_c is proportional to judged probability: $I_c \approx$ P(H|D). Figure 1 illustrates the conditional likelihood judgment process in a flow chart for the judgment P(democrat|wealthy).

MDM Applied to Overconfidence.

Figure 2 presents a conceptual model of the probability estimation process. For example, imagine that participants are presented with the general knowledge question: "Which city is larger: Los Angeles or Sacramento?" There are two ways that this

judgment can be made: 1) retrieval of a surrogate cue that points to an answer (e.g., population statistics, number of sports teams), or 2) infer on the basis of familiarity or echo intensity (cf. Yonelinas, 1994). MDM is used to model this second type of judgment process and is not applied to situations where participants can infer the answer on the basis of specific knowledge, or where surrogate cues are used to infer the answer (cf. Gigerenzer et al., 1991).[2]

If an answer is recalled from memory, it is assumed that the participant responds with the recalled answer and gives a confidence judgment of 100%. Thus, if one can recall the populations of Los Angeles and Sacramento, or can recall a cue that points to Los Angeles as larger than Sacramento, then it is assumed that the participant will choose Los Angeles with 100% confidence. However, if an answer cannot be recalled (with 100% certainty), it is assumed that participants probe memory with all alternative answers and assess the relative echo intensity or familiarity. For the above question, participants are assumed to probe memory with Los Angeles and Sacramento. If the relative echo intensity values are equivalent, then the participant is assumed to choose at random and respond with 50% certainty. If one alternative returns a relatively higher echo intensity, the decision rule in MDM is to choose the alternative with the highest echo intensity. Confidence is assumed to be proportional to the relative echo intensity or conditional echo intensity of the x-alternative answers. For

[2] Although this might seem to somewhat limit MDM's applicability, previous research indicates that choice behavior often can be attributed to recognition processes (Gigerenzer & Goldstein, 1996). For example, in one study, Goldstein (1998) found that an average of 93% of participant's choices were made on the basis of recognition memory. This leaves little left to be explained by the use of surrogate cues.

the above question, most people will choose Los Angeles because it is the most familiar. Confidence is assumed to be given by: confidence = $I_{\text{Los Angeles}}/(I_{\text{Los Angeles}}+I_{\text{Sacramento}})$. Note that for probabilistic tasks where the chosen answer is correct only with some probability, the decision maker will not be able to recall the "correct" answer with 100% certainty. Thus, in these situations, the decision maker will, by necessity, choose the answer by assessing the relative familiarity of the x-alternative answers.

There are three factors that can be modeled by MDM and that affect the model's overconfidence predictions: 1) the ecological cue structure or probability, 2) experience, and 3) encoding. All three factors can be seen as arising from the interaction between the decision maker and the environment and therefore can be modeled by combining the Brunswikian assumptions of ecological structure with MDM's assumptions of memory representation. These three factors will be discussed within the context of the ecological models and then related to the error model in the sections to follow.

Ecological Models

Ecological cue structure. The first factor that can be modeled by MDM is the ecological cue structure. For present purposes, the ecological cue structure is defined as the *ratio of traces* for each answer. Ecological cue probability can be accommodated naturally by any instance-based model if it is assumed that people's past experiences are stored in a manner similar to how they are encountered in the environment. This frequentistic memory representation enables cue probabilities to be inferred from

memory by comparing relative frequencies. Let f(A) and f(B) represent the frequency of traces in memory corresponding to two mutually exclusive events, A and B. Assume further that the decision maker's task is to make a comparative judgment of some sort, such as "which city is larger, A or B?" The ecological cue probability that A is the larger city is assumed to be given by the relative frequency (i.e., ratio) of A to B traces stored in memory: f(A)/[f(A)+f(B)].

Returning to the earlier example, Los Angeles is a more famous city than Sacramento and people probably have more traces in memory corresponding to Los Angeles than to Sacramento. The ratio of traces stored in memory for each alternative answer for this question (i.e., Los Angeles and Sacramento) directly affects the relative echo intensity and how often participants will choose one answer versus the other as correct. Assume that the ratio of Los Angeles to Sacramento traces is 5 to 1 (5 traces of Los Angeles for every 1 trace of Sacramento), then the ecological cue probability is given by f(Los Angeles) / [f(Los Angeles)+f(Sacramento)] = 5 / (5+1) = .83.[3]

For conditional probability judgments, the ecological cue probability is assumed to equal the objective *conditional* probability. For example, imagine that the task is to judge the probability of $disease_1$ given a symptom (e.g., P($disease_1$|symptom)). The ecological cue probability can be defined as the relative frequency of $disease_1$ associated with the given symptom or cue: f(disease1|symptom)/f(symptom). Thus, the ecological cue probability is a valid Bayesian posterior probability. Of course, the use of a symptom as a cue is just one example; the concept can be generalized to account

for any other type of conditional probability judgment that is generated on the basis of a cue or multiple cues. For example, one can think of estimating the probability that a city is a capital given that it has national monuments (e.g., P[capital|monuments]), or the probability of a capital given monuments and a large population (e.g., P[capital|monuments$\cap$large population]).

Notice that for conditional probability judgments, the ecological reference class is defined by the condition in the conditional judgment (e.g., by monuments in the P[capital|monuments] judgment). Note also that this assumption is built into MDM's conditional memory retrieval mechanism as seen in Eq. 4.

Experience. MDM accounts for two aspects of experience. The first aspect of experience that can be accommodated by MDM is that having to do with cognitive adjustment (i.e., how information is sampled from, or is encountered in, the environment; Brunswik, 1964; Gigerenzer et al., 1991; Juslin et al., 1997). Even though the objective ecological cue probabilities for Los Angeles and Sacramento might be 5:1 (i.e., Los Angeles is discussed 5 times more often than Sacramento), the extent to which our memory corresponds to this ratio will depend on how much experience we have. To the extent that we have experience in an ecological reference class, the memory representation of the cues in the environment will approach the true ecological cue probabilities. With little experience, the memory representation of the environmental cues might contain considerable external sampling error (i.e., we may, by chance, have 7 traces of Los Angeles and only 1 trace of Sacramento). However, as

[3] Note that this ratio (and ecological cue probability) can remain constant, yet the absolute frequencies change (e.g., 5:1 = 50:10). As will be elaborated shortly, MDM actually makes quite different

one gains experience in the environment, external sampling error will decrease, and the memory representation will become a more accurate representation of the ecological cue probabilities (a result of the law of large numbers).

The second aspect of experience that can be accommodated by MDM is trace frequency. Even assuming that the internal memory representation accurately reflects the external ecological cue structure, experience can have an indirect, yet profound, impact on judgment. MDM accounts for the effect of experience by assuming that the frequency of task-relevant traces stored in memory increases with experience. Assume a two-alternative general knowledge question. As one gains experience in an ecological reference class, the frequency of traces that are similar to one another will increase. For example, mere experience can lead to an increase in the number of traces in memory corresponding to Los Angeles and to Sacramento. The effect of increasing the number of traces that have a high degree of inter-trace similarity is improved calibration: Overconfidence will decrease as the frequency of similar task-relevant memory traces increases – this will be referred to as the *trace-frequency prediction.*

Previous research in decision making has found that experts in several domains are often better calibrated than novices (e.g., bridge, Keren, 1987; weather forecasting, Murphy & Winkler, 1977; accounting, Mladenovic & Simnett, 1994; Tomassini, Solomon, Romney, & Krogstad, 1982). These results are consistent with MDM's trace-frequency prediction if it is assumed that these experts had more task-relevant instances stored in memory.

---

predictions in these two cases.

Encoding or information loss. A third characteristic of the natural environment that can be modeled by MDM is how well information is encoded in memory. Although, encoding is a psychological variable, it can be viewed as being driven by environmental factors. Environmental events often receive differing degrees of attention: Interesting and important stimuli tend to draw attention while uninteresting or unimportant stimuli tend not to receive much attention. Assuming that attention is necessary for encoding (Boronat & Logan, 1997; Craik, Govoni, Naveh-Benjamin, & Anderson, 1996), some environmental stimuli will receive a better quality of encoding than other, less interesting or less important, stimuli (e.g., Gronlund, Ohrt, Dougherty, Perry, & Manning, 1998).

In MDM, the quality of encoding is modeled by the encoding parameter, L, which determines how well information is stored in memory. The net effect of improved encoding quality in MDM is a reduction in overconfidence. In particular, MDM predicts that overconfidence will decrease under conditions that lead to high levels (better quality) of encoding – the *encoding quality prediction.*

At least one study has shown a relationship between the quality of encoding and calibration. Juslin, Olssen, and Winman (1996) examined how well participants' judgments were calibrated for stimuli that were either central to the focus of attention or peripheral to the focus of attention. Overconfidence was lower for stimuli that were central to the focus of attention. Presumably this was due to a difference in how well the different stimuli were encoded.

The above discussion illustrated how MDM accommodates several aspects of the ecological approach. Importantly, the assumptions built into MDM as a consequence of this integration have implications both for MDM's overconfidence predictions and for the understanding of the error models. In the next section I discuss how experience and encoding affect the random error produced in the memory retrieval process.

Error models.

Erev et al. (1994) showed that overconfidence could be accounted for by a model that assumed that true judgments were perturbed by random error. This random error was assumed to be generated in the response-selection phase of judgment: The result of momentary fluctuations in the response process (Thurstone, 1927).

Rather than arising from response processes per se, random variation in MDM arises from how information is retrieved from memory, in particular, from the computations on the vectors stored in memory. One property of the Erev et al. model is that overconfidence decreases as error variance decreases. Interestingly, the two factors that lead to better calibration in MDM – experience and encoding – also lead to the reduction of error variance in the model. Error variance decreases as encoding quality increases, the result of computations on the vectors, and as experience increases, a result of the law of large numbers. Improving encoding quality decreases variance because there are fewer 0's present in the memory trace, the more 0's present in the trace, the less similar the trace will be to the probe vector. In contrast, increasing experience decreases variability because there is an increase in the frequency of traces

in memory that have a high degree of similarity to one another. Thus, the reduction in error variance in this case is a direct result of the law of large numbers (increasing N naturally reduces variability).[4]

The net result of increased experience and improved encoding quality is the reduction of error variance in memory retrieval, which in turn leads the model to choose the alternative with the highest ecological cue probability more frequently. Consider two cases as examples: One in which the participant has 5 traces of Los Angeles and 1 trace of Sacramento, and the second in which the participant has 500 traces of Los Angeles and 100 traces of Sacramento. The variability associated with the first case will necessarily be higher than that associated with the second case because there are only 6 total traces processed at memory retrieval in the former, but 600 traces processed in the later. This variability in turn affects the probability that participants will choose the alternative with the highest ecological cue probability: The higher this error variance is, the greater the chance that the echo intensity will be higher for Sacramento. This is true even though the trace ratios (and by extension, the ecological cue probabilities) are equivalent. Thus, increasing the number of traces, while keeping the ecological cue probability constant, should lead to an increase in proportion correct, but leave unaffected the mean probability judgment.

Encoding also affects the amount of error variance produced at retrieval. As encoding quality increases, the error resulting from the vector computations decreases (this is a consequence of decreasing the number of zeros in the memory vectors).

---

[4] The decrease in variability with the increase in the number of traces is contingent on the traces having a high degree of similarity to one another. Adding traces to memory that are orthogonal (i.e.,

Again, this reduction in error variance leads the model to choose the alternative with the highest ecological cue probability more often. Thus, assuming that encoding is unbiased, increasing encoding quality should lead to an increase in proportion correct without affecting the mean probability judgment.

The above discussion illustrates that MDM is consistent with the Thurstonian account of random variation. However, in contrast to the Thurstonian account, MDM specifies the locus of the random variation and the factors that affect it. More importantly, MDM's account of overconfidence shows that the Brunswikian and Thurstonian models can be accommodated by a single memory-processes model. By instantiating the factor of experience in the context of MDM, it was shown that the same factor that the Brunswikians argued was necessary for good calibration (experience within the ecological reference class from which the questions are drawn) also leads to less variation at memory retrieval. In addition, the present analysis makes the novel prediction that encoding processes are fundamental to good calibration. I now turn to illustrating these predictions through simulations.

Simulations and Model Predictions

Four simulations using MDM were done to simulate the effects of experience and encoding on the calibration of probability judgments. The purpose of these simulations was to demonstrate MDM's encoding quality and experience predictions. The experiments presented later will provide tests of these predictions.

For each simulation, 1000 participants were simulated and S_c was arbitrarily set to .60. The first set of simulations examined MDM's encoding quality predictions. In

---

dissimilar) to the target actually increases variability (see Gronlund & Elam, 1994).

these simulations, the effect of encoding was modeled by varying the encoding parameter: L was set to .35 in the poor-encoding condition and .55 in the good-encoding condition. Eighty instances were stored in memory for each simulated participant. The relevant trace frequencies and the objective ecological cue probabilities for each H|D combination are presented in Table 1 (the numbers not in parentheses).

The second set of simulations examined MDM's experience (i.e., trace frequency) predictions. This was done by varying the frequency of traces stored in memory across two levels. In the high-experience condition, a total of 240 traces were stored in memory and in the low-experience condition a total of 80 traces were stored in memory. The exact frequency of traces corresponding to each H|D combination are presented in Table 1. For example, in the low-experience condition, there are 9 traces of H_1 and 1 trace of H_2 for D_1 (the corresponding frequencies for the high-experience condition were 27 and 3). $L = .35$ and $S_c = .60$ for both simulations.

All four of the above simulations model a decision task in which the D's are probabilistically related to the H's. This type of task is comparable to a disease-diagnosis task in which the symptoms (the D's) are probabilistically related to the diseases (the H's). A simulation of a probabilistic task was chosen because it represents the more general case where the outcome variable is probabilistically related to the predictor variables (for comparison, in the general knowledge paradigm, the outcome variable can be determined with certainty) and because Experiments 1 and 2 used this type of probabilistic task.

Figure 3 presents the results of the simulations where encoding quality is varied (top panel) and where experience is varied (bottom panel). As can be seen, there was a clear effect of experience and encoding on overconfidence: Overconfidence decreased as experience increased and as encoding quality (as modeled by L) increased. The effect of experience on overconfidence is the direct result of increasing the frequency of similar traces stored in memory. As the frequency of similar traces increases, the error variance associated with retrieval decreases and the model chooses the alternative with the highest ecological cue probability more often (i.e., proportion correct increases). Increasing encoding quality has a similar effect on overconfidence and it too is the result of a decrease in the random error. However, in this case, the random error is the result of individual vector computations. As encoding quality increases, the error produced by the individual vector computations decreases.

A second, and less obvious finding from these simulations is that MDM's reduction in overconfidence is due mostly to the increase in proportion correct, as opposed to decreases in the predicted probability (i.e., proportion correct increased, but the mean predicted probabilities remained constant). This is most obviously seen by comparing the two lines in each of the two graphs. Notice that the mean predicted probability is unchanged even when encoding quality improves and when experience increases.[5] This suggests that, at least for probabilistic tasks where the trace ratios are constant, MDM's predictions are due to an increase in the proportion correct and not better attunement of probability judgments.

Experiment 1

The purpose of Experiment 1 was to test MDM's prediction that overconfidence decreases as encoding quality improves. Past research on the overconfidence phenomenon has used the general knowledge paradigm. The general knowledge paradigm is the limiting case of MDM's account of the overconfidence phenomenon as it represents the case where the outcome variable is deterministic: Either the decision maker is correct or incorrect. The more general paradigm for studying overconfidence is a probabilistic paradigm in which the outcome variable is probabilistically related to the predictor variable.

Experiments 1 and 2 used a probabilistic decision task in which participants studied a population of fictitious patients who were suffering from one of two diseases. The two diseases were probabilistically related to a set of 8 mutually exclusive and exhaustive symptoms. The participant's task in Experiments 1 and 2 was to first study a population of patients in which both the symptom and the disease are known, and second, to diagnose a new population of patients characterized by the same symptoms but in which the diseases are unknown. The major advantage this task has over the general knowledge paradigm is that the frequency of traces, and therefore, the ecological cue probabilities, can be manipulated. In contrast, in the general knowledge paradigm, it is impossible to know a priori how much exposure participants have with a particular set of general knowledge questions.

---

[5] This is due to the fact that the ratio of traces remains constant across all conditions. Thus, even though the variance is higher when encoding quality is poor and when experience is low, the mean predicted probability judgments remains the same.

Method

Participants

Participants were 79 undergraduate students enrolled in lower-level psychology courses at the University of Oklahoma. Participants received partial course credit in return for their participation.

Design and Procedure

The experimental design for Experiment 1 was a 2-group between-subjects design with encoding quality (good, poor) as the independent variable. The experimental task was a disease-diagnosis task consisting of a study phase and a testing phase. During the study phase of the experiment, participants studied a population of fictitious patients, each described by a single symptom and the associated disease. Participants were told that they would be studying a number of such patients that had already been diagnosed and that they were to learn which symptoms went with which disease. They were also informed that all eight symptoms had the possibility of occurring with either disease.

Table 1 presents the disease-symptom combinations used in the study phase of Experiment 1, the actual frequency with which each disease-symptom occurred (i.e., the trace ratios), and the objective ecological cue probabilities. Participants in both the good and poor encoding quality condition studied 80 fictitious patients (the numbers in parentheses correspond to the frequency of traces used in the high-experience condition of Experiment 2).

Participants in both the good and the poor encoding conditions were given two different color drawings of faces, each labeled with one of the two diseases (metalytis, zymosis) (cf. McKenzie, 1998). Participants in the good-encoding condition were told that the experiment was testing a new technique for improving memory performance that involved mental visualization. They were instructed to visualize mentally each symptom as it would appear on the patient. For example, if the symptom "red rash" occurred with metalytis, the participant was to mentally picture a "red rash" occurring on the face of the metalytis patient. Participants in the poor-encoding condition were told only that they were to study each fictitious patient as it appeared and that they would be asked several questions later about what they had learned.

The fictitious patients were presented on the computer screen one at a time and in random order. Although participants were self-paced, the software ensured that each patient was presented on the computer screen for no less than 2.0 s. A recognition test was given following randomly chosen patients during the study phase to ensure that participants were motivated to remember the diseases and symptoms. The recognition test required participants to identify either the disease or the symptom of the patient just presented. The probability that a given patient was followed by a recognition test was .20.

Following the study phase was a brief (5 minute) intervening task in which participants completed two individual difference scales (need for cognition [Caccioppo & Petty, 1982], and tolerance for ambiguity [MacDonald, 1970]). These tasks served as distracter tasks and were not part of the experimental design.

In the testing phase of the experiment, participants were presented with a symptom (e.g., pallor) and had to choose which disease they thought was most likely and then estimate the conditional probability P(Disease | Symptom). The probability judgments were made on a half-range response scale (i.e., .50 to 1.0) by entering their response into the computer. Participants were instructed to be as accurate as possible when making their probability judgments and to "feel free to use any number between .50 and 1.0 that most accurately reflected the true probability." Participants were instructed to base their choice of disease and their probability judgment on their memory for how often the symptom occurred with both diseases. Each participant made 6 sets of judgments (diagnosis and probability estimate) for each of the 8 disease - symptom combinations resulting in 48 total sets of judgments.

Dependent measures of accuracy

The accuracy of probability judgments can be assessed using proper scoring rules. One such scoring rule is the probability score (Brier, 1950; Yates, 1991). The probability score is a measure of the accuracy of one's probability judgments and is described by the equation:

$$\text{probability score} = \frac{1}{N}\sum_{i=1}^{N}(r_i - c_i)^2 \qquad (5)$$

where N is the total number of probability assessments, r_i is the i^{th} probability judgment, and c_i is the outcome index for the i^{th} judgment (1 if correct and 0 if incorrect). Therefore, the probability score is the average of the squared differences between each probability judgment (r_i) and the outcome index (c_i). Notice that lower probability scores indicate better accuracy, with a score of 0.0 corresponding to perfect accuracy

(in which case the judge assigns a probability of 1.0 and is correct on all judgments) and a score of 1.0 corresponding to perfect *inaccuracy* (in which case the judge assigns a probability of 1.0 and is *incorrect* on all judgments). Several decompositions of the probability score have been proposed, but those proposed by Yates (1982) and Murphy (1973) are the most popular. Yates's (1982; 1991) decomposition was used in the present experiments.

Yates' Decomposition of the Probability Score

Yates proposed an additive decomposition consisting of 4 parts: probability score = knowledge + calibration in-the-large + knowledge(slope)(slope-2) + scatter. Each component of the decomposition reveals different aspects of the quality of people's judgments

Knowledge describes the overall accuracy of the judge and is defined by the equation:

$$\text{Knowledge} = C(1-C) \qquad (6),$$

where C is the proportion of correct responses across all judgments. Again, assuming that the participant performs better than chance, lower knowledge scores are better, with 0.0 indicating perfect knowledge (i.e., the proportion of correct responses is 1.0). In a two-alternative choice task (as were used in the present experiments), a knowledge score of .25 indicates chance performance (i.e., the proportion correct is .50).

Calibration in-the-large is a measure of the accuracy of one's average probability estimate compared to his or her mean probability judgment and is defined as:

$$\text{Calibration in-the-large} = (C - R)^2 \quad (7),$$

where R is the mean of all probability judgments and C is defined as previously stated. Calibration in-the-large is best if close to 0 and poor as it deviates from 0. Notice that calibration in-the-large does not indicate whether the judge is over- or underconfident.

The third component of Yates' decomposition is slope. Slope is described as the difference between the mean probability estimate conditionalized on correct responses and the mean probability estimate conditionalized on incorrect responses:

$$\text{slope} = R_1 - R_0 \quad (8)$$

where R_1 is the mean probability judgment for correct responses and R_0 is the mean probability for incorrect judgment. Ideally, the judge should assign higher probability judgments when he or she is correct, and lower probability judgment when he or she is incorrect. Thus, a higher slope indicates better ability to discriminate between correct and incorrect judgments.

The final component of Yates' decomposition is scatter. Scatter is a measure of variability in the judges' probability judgments conditional on whether his or her responses are correct or incorrect, and is defined as:

$$\text{scatter} = \frac{N_1 Var(r_1) + N_0 Var(r_0)}{N_1 + N_0} \quad (9),$$

where N_1 and N_0 are the frequency of correct and incorrect responses respectively, and $var(r_1)$ and $var(r_0)$ are the variances conditional on correct and incorrect responses respectively. Thus, scatter is a weighted mean of conditional variances (Yates, 1991).

A final measure typically used to measure accuracy is the bias or over/underconfidence score. This is simply the difference between the overall mean probability judgment R and the proportion correct C: Bias = R-C. Overconfidence obtains when the bias score is positive and underconfidence obtains when the score is negative.

Results and Discussion

Figure 4 presents group calibration curves for the good- and poor-encoding conditions in Experiment 1. As can be seen, there was a clear effect of the encoding manipulation. Notice that the plot for the good-encoding condition is noticeably closer to the identity line than the plot for the poor-encoding condition. This pattern of data is consistent with MDM's account of overconfidence as shown in the top half of Figure 3, namely that overconfidence should decrease as encoding quality improves.

Eight one-way ANOVA's were done on the probability scores, Yates' decomposition of the probability score, the mean proportion correct, and the mean probability estimates. An alpha of .05 was used for all significance tests unless otherwise noted.

Table 2 presents the various measures of accuracy along with the F-statistics and the effect sizes calculated using Cohen's *d*. There was a significant main effect of encoding on the overall probability score, knowledge, and bias, with participants in the good-encoding condition performing significantly better (lower scores) on all three of these measures. In addition, good-encoding participants had marginally significantly better (higher) slopes. In each of these cases, the effect sizes were moderately high.

Importantly, the above results were driven primarily by the improvement in proportion correct: The mean probability score, knowledge, and bias decreased as a result of the increase in proportion correct. An important finding in this experiment was that there was no effect of encoding on participant's probability judgments. Instead, the decrease in overconfidence (bias) was due solely to the increase in the proportion correct (i.e., participants did not adjust their probability judgments to be in line with their proportion correct). This is consistent with MDM's account: That mean probability judgment should remain the same, but proportion correct should increase as encoding quality improves.

The findings of Experiment 1 supported MDM's encoding-quality predictions. Namely, overconfidence decreased as encoding quality increased. Experiment 2 tested MDM's trace-frequency prediction.

Experiment 2

The purpose of Experiment 2 was to test MDM's prediction that overconfidence decreases as trace frequency increases. Participants in the high-experience condition studied 3 times the number of fictitious patients in the study phase than were studied by participants in the low-experience condition. All participants in Experiment 2 received the poor-encoding instructions and were not provided drawings of faces on which the symptoms could be mentally visualized.

Method

Participants

Participants were 58 undergraduate students enrolled in psychology courses. Participants received partial course credit in return for their participation.

Design and Procedure

The experimental design was a two-group between-subjects design with experience (80 vs. 240 study trials) as the independent variable.

The basic experimental task was identical to that of Experiment 1 with the following exceptions. First, rather than manipulating encoding, all participants received the poor-encoding instructions from Experiment 1. Experience was manipulated by varying the frequency of patients presented during the study phase of the experiment. Participants in the low-experience condition were presented with 80 fictitious patients and participants in the high-experience condition were presented with 240 fictitious patients. Table 1 presents the actual frequency that each disease-symptom occurred (the trace ratios) for both the 80 and 240 study trials conditions (the frequencies for the 240 condition are in parentheses).

Following the study phase, participants were given a brief (5 minute) intermittent task that required them to solve 50 mathematical problems. This task was done to reduce recency effects and was not part of the experimental design.

Results and Discussion

Due to the high degree of between-subjects variability, two outliers were eliminated from the sample using the trimmed mean method on the bias score (e.g.,

participants whose bias scores were $\pm 2.5\sigma$ from the mean were deleted). Both outliers were from the low-experience condition.

Figure 5 presents group calibration curves for the high- and low-experience conditions. There was a clear effect of experience, with the high-experience condition showing much less overconfidence than the low-experience condition. This is shown clearly by the fact that the plot for the high-experience condition is closer to the identity line.

Analyses of variance (ANOVA) were performed on the probability score, Yates's decomposition, bias, mean proportion correct, and the mean probability judgment. Table 3 shows the results of these analyses along with the effect-size calculations. There was a significant effect of experience on the probability score, the knowledge score, and bias, with the high-experience condition showing better performance (lower scores) on all three measures. In addition, participants in the high-experience condition had better (higher) slope scores, though this was significant only at the .13 level. Similar to Experiment 1, the effect sizes for the probability score, bias, and knowledge were in the high range ($d = .70$ to $.81$) and the effect size for slope was in the medium range ($d = .42$). Finally, consistent with MDM's predictions, mean confidence did not differ between the high and low experience conditions. Thus, the reduction in overconfidence was again due solely to the increase in proportion correct.

The results of Experiment 2 support MDM's prediction that experience leads to decreased overconfidence and that this decrease is due to an increase in proportion correct. However, one question that remains unanswered by Experiments 1 and 2 is

whether the encoding-quality and trace-frequency predictions extend to the general knowledge paradigm. This question was answered in Experiment 3.

Experiment 3

The purpose of Experiment 3 was to extend the trace-frequency and encoding manipulations to the general knowledge task typically used in overconfidence research. Past research using the general knowledge paradigm has simply presented participants with a series of two-alternative force-choice general-knowledge questions. Each question is answered and confidence is assessed on a half-range (.5 to 1.0) confidence scale.

The methodology used in the present research was somewhat different, and consisted of two phases. In the first phase of the experiment, participants studied a subset of the general knowledge questions with their associated answers. A third of the total set of question/answer pairs appeared three times during the study phase and a third appeared only once during the study phase. The remaining one-third of the questions did not appear during the study phase. Thus, trace-frequency was manipulated across three levels: 3, 1, and 0 presentations. Encoding was manipulated by having half of the participants engage in mental visualization to improve their memory for the questions/answers. The remaining participants were told to read the questions and answers as quickly as possible.

The general knowledge questions selected for this experiment were selected to be of high difficulty. This was done to minimize the effect of participants' prior knowledge. If participants could rely on their prior knowledge to answer the questions,

the effectiveness of the trace-frequency and encoding-quality manipulations would be minimal. The goal of the experiment was to provide participants with the relevant "experience" in the study phase of the experiment where trace frequency and encoding quality could be controlled.

Overconfidence should decrease both as encoding quality increased and as trace frequency increased for the questions studied during the study phase. In contrast to the previous studies where the ratio of traces was held constant across encoding and experience conditions, the ratios necessarily varied (because only the correct answer to each question was studied). Thus, while encoding and experience affected only proportion correct in the previous studies, they should affect both proportion correct and mean confidence in this experiment.

Method

Participants

Participants were 50 undergraduate students enrolled in lower-level psychology courses. Participants received partial course credit in return for participation.

Design and Procedure

The design of this experiment was a 2 (encoding) x 3 (trace frequency) mixed factorial with encoding quality (good, poor) as the between factor and trace frequency (3, 1, 0 presentations) as the within factor.

Materials. One-hundred sixty-two general knowledge questions were selected as stimuli. The general knowledge questions were specifically selected to be of very high difficulty, to minimize any effect of participants' prior knowledge of the questions.

Procedure. The experiment consisted of two phases. In the study phase, participants studied 108 different general knowledge questions with the answers provided (e.g., "Other than the sun, which star is closer to earth? Proxima Centauri"). 54 occurred 3 times during the study phase and 54 occurred only once, for a total of 216 items studied during the study phase (54 x 3 + 54 x 1 = 216). Fifty-four additional questions were not presented in the study phase (0 presentations) but were presented in the testing phase. Questions were randomly assigned to each of the presentation conditions and were the same for all participants (i.e., questions were not randomly assigned for each participant). The questions were presented by computer one at a time in a random order, and participants were run individually on computers.

Encoding quality was manipulated by having participants engage in one of two orienting tasks. Participants in the poor-encoding condition were told that the experiment was measuring the effects of familiarity on how quickly they could read general knowledge questions. They were told to read the questions and answers as quickly as possible as they were being timed on how long it took them to read each question/answer.

Participants in the good-encoding condition were told that the goal of the experiment was to test the effectiveness of mental imagery in remembering general knowledge facts. Participants in the good-encoding condition were instructed to form a mental image containing the information in the question and the answer for each item that was presented during the study phase. If they were unable to form a mental image, they were instructed to associate a word in the question with the given answer. It was

emphasized that they would be tested on their memory for the answers to the questions at a later date and that their performance on the memory test would depend on how well they remembered the questions.

The testing phase of the experiment took place four days later. Participants were told that they would be answering a series of general knowledge questions and that many, but not all, of the questions had appeared in the study session. The general knowledge test consisted of a two-alternative forced-choice test. Participants were presented with a question along with two alternative answers (e.g., "Other than the sun, which star is closer to earth? A) Proxima Centauri B) Barnard's Star"). After answering each question, participants rated their confidence in their answer by adjusting a tick mark on a number line anchored with "50% certain" at one end and "100% certain" at the other end. Each participant answered a total of 162 general knowledge questions: 54 had occurred 3 times (3 presentations) in the study phase, 54 had occurred 1 time (1 presentation) in the study phase, and 54 of the questions were new (0 presentations).

Results and Discussion

Figure 6 illustrates the calibration curves for the good- (top panel) and poor-encoding (bottom panel) conditions for each of the three trace-frequency conditions. First, note that for both encoding conditions, the plot for the 0 presentations (dashed line) condition is well below the identity line. This illustrates that both the good- and poor-encoding groups showed a large degree of overconfidence for the questions that

did not appear in the study phase. Thus, in the absence of prior experience, participants showed the typical overconfidence result.

The two solid lines in Figure 6 plot the calibration curves for the questions that occurred once (open circles) and the questions that occurred 3 times (filled circles) during the study session. Notice that the plot for the 3-presentations condition is closer to the identity line than the plot for the 1-presentation condition, showing that participants were less overconfident for questions with which they had more prior experience. Finally, notice that the 3- and 1-presentation calibration plots are generally closer to the identity line in the good-encoding condition. This shows that overconfidence decreased when encoding quality increased.

Eight two-way repeated measures ANOVA's were done to test the effects of experience and trace frequency on the probability score and Yates's decomposition of the probability score.

Recall that main effects of encoding and trace frequency were predicted, as was an encoding x trace-frequency interaction. Specifically, the good-encoding condition was predicted to be less overconfident than the poor-encoding condition for the 3- and 1- presentation conditions, but not for the 0-presentation condition. In addition, participants were predicted to be less overconfident in the 3-presentation condition than in the 1-presentation condition, and less overconfident in the 1-presentation condition than in the 0-presentation condition.

Table 4 shows the mean probability scores and Yates's decomposition of the probability score for each level of encoding and trace-frequency. As expected, there

were significant main effects of encoding on the probability score (F[1,48] = 32.52, Mse = .008), bias (F[1,48] = 11.29, Mse = .021), knowledge (F[1,48] = 49.63, Mse = .002), slope (F[1,48] = 4.25, Mse = .006), calibration-in-the-large (F[1,48] = 7.88, Mse = .002), and scatter (F[1,48] = 4.33, Mse = .0001). In all cases, the good-encoding condition performed significantly better than the poor-encoding condition.

There were also significant main effects of trace-frequency on the probability score (F[2,47] = 195.37, Mse = .002), bias (F[2,47] = 67.66, Mse = .005), knowledge (F[2,47] = 207.76, Mse = .001), slope (F[2,47] = 37.49, Mse = .004), calibration-in-the-large (F[2,47] = 32.95, Mse = .001), and scatter (F[2,48] = 20.97, Mse = .00005). Generally, participants performed better in the 3-presentation condition than in the 1-presentation condition, but performed better in the 1-presentation condition than in the 0-presentation condition.

Tests of the encoding x trace-frequency interaction revealed significant interactions for the probability score (F[2, 47] = 8.35, Mse =.002), knowledge (F[2,47] = 25.07, Mse =.001), slope (F[2, 47] = 3.38, Mse =.004), and scatter (F[2,47] = 6.66, Mse = .00005) and a marginally significant interaction for bias (F[2, 47] = 2.59, Mse =.005, p =.08). The interaction for calibration-in-the-large-failed to reach significance (F[2,47] = 1.93, Mse =.001, p = .15).

Univariate tests of each trace-frequency condition, and inspection of the means in Table 4, revealed that the interactions were generally due to small or negligible differences between the good- and poor-encoding conditions for the 0-presentation condition relative to the 1- and 3-presentation conditions. This pattern generally held

true for the probability score (0-presentation $F[1,48]=2.69$, $Mse = .004$, $p = .10$, 1-presentation $F[1,48]=33.97$, $Mse = .005$, 3-presentation $F[1,48] = 37.33$, $Mse = .004$), bias (0-presentation $F[1,48]=3.35$, $Mse=.015$, $p = .07$, 1-presentation $F[1,48]=14.35$, $Mse=.011$, 3-presentation $F[1,48] = 8.39$, $Mse=.006$), and knowledge (0-presentation $F[1,48] = 0.64$, $Mse = .000$, $p = .42$, 1-presentation $F[1,48] = 35.18$, $Mse = .002$, 3-presentation $F[1,48] = 48.60$, $Mse = .002$). Although the encoding x trace frequency interaction failed to reach significance for calibration-in-the-large, the univariate tests for each level of trace frequency revealed the predicted pattern. There was no significant effect of encoding on calibration-in-the-large for the 0-presentation condition ($F[1,48] = 1.91$, $Mse = .003$, $p = .17$) but significant effects for the 1- and 3-presentation condition ($F[1,48] = 11.46$, $Mse = .001$ and $F[1,48] = 7.27$, $Mse = .0004$ respectively). The lone exceptions to the predicted pattern were for slope and scatter. Although the interaction for slope was significant, it was apparently due solely to a difference between the good- and poor-encoding conditions for the 1-presentation condition, as there was no effect of encoding quality in the 3-presentations condition (0-presentation ($F[1,48] = 1.34$, $Mse = .001$, $p = .25$, 1-presentation $F[1,48] = 9.90$, $Mse = .004$, and 3-presentation $F[1,48] = 0.01$, $Mse = .008$, $p = .91$). For scatter, there were no differences between the good- and poor-encoding conditions for the 0- and 1-presentation condition (0-presentations $F[1,48] = 0.15$, $Mse = .0001$, 1-presentation $F[1,48] = 2.08$, $Mse = .00009$). However, participants in the good-encoding group had significantly less scatter in the 3-presentation condition ($F[1,48] = 27.01$, $Mse = .00004$).

The above results generally support MDM's prediction that calibration should improve as encoding quality increases and as trace-frequency increases. For virtually all of the measures of accuracy, participants performed better as experience increased (from 0 to 3 presentations of the question/answer) and when encoding quality was good.

There were negligible differences between the good- and poor-encoding conditions for the 0-presentation condition for all the dependent measures. This is important for two reasons. First, it indicates that there were no a priori differences between participants in the good- and poor-encoding conditions: Both groups were equally poorly calibrated for questions that did not appear in the study phase of the experiment. Second, and more important, it suggests that the improvement in accuracy resulting from studying the question/answers was not due to metacognitive processes, but instead was due to the retrievability of information from memory. Had there been systematic differences between the good- and poor-encoding conditions for mean confidence in the 0-presentation condition, it would have pointed to metacognitive processes as a moderator of overconfidence.

It is easy to imagine how such metacognitive processes might operate in the context of this experiment. For example, one way to estimate one's confidence in an answer is to base the judgment on the difference between the familiarity of the currently-being-judged answer and the average familiarity of answers to prior questions. (In contrast, the assumption made by MDM is that confidence is based on the relative familiarity of the two-alternative answers posed in the general knowledge

question.) For the 0-presentation condition, the difference between the familiarity of the chosen answer and previously answered questions would be smaller in the poor-encoding quality condition than in the good-encoding quality condition. This is because the average familiarity of the 1- and 3- presentation conditions for the good-encoding quality condition should be higher than the average familiarity of the 1- and 3-presentation conditions for the poor-encoding condition, relative to the 0-presentation questions. Thus, if participants were basing their confidence judgments on the relative familiarity of *past* answers, mean confidence would be lower in the good-encoding 0-presentation condition than in the poor-encoding 0-presentation condition. Although mean confidence was slightly lower in the good-encoding condition, it did not approach significance ($p = .51$). The overall improved calibration for the high-encoding condition and for the 1- and 3- presentation conditions cannot be attributed to this metacognitive process.

Conclusions

The present research has shown that overconfidence is intimately tied to how well information can be retrieved from memory and that it decreases as experience and encoding quality increase. Experiments 1 and 2 demonstrated the effect of encoding and experience on overconfidence using a probabilistic task. Overconfidence was lower when encoding quality was good (Experiment 1) and when experience was high (Experiment 2). Consistent with MDM, the reduction of overconfidence in both experiments was due primarily to an increase in proportion correct and not to changes in confidence judgments. Experiment 3 replicated the findings of Experiments 1 and 2

using the general knowledge paradigm. In sum, all three experiments support MDM's predictions that the factors of encoding and experience are fundamental to overconfidence.

Implications for Models of Overconfidence

Keren (1991) argued that much of the research on the overconfidence phenomenon lacked a coherent theoretical framework. However, since the publication of his article, two prominent and seemingly different theories have emerged. One of these models, the ecological or Brunswikian model (Gigerenzer et al., 1991), placed the locus of overconfidence in the environment: The result of a biased selection of general knowledge questions on the part of the experimenter (Juslin, 1994; Juslin et al., 1997). The other of these models, the random error or Thurstonian model (Erev et al., 1994), placed the locus of overconfidence as internal to the participant: The result of noisy psychological processes. Indeed, both of these models have made important contributions to the literature, and both have changed the way in which the overconfidence phenomenon is conceptualized.

Arguably the most important question generated by these models is whether overconfidence is real or simply an artifact (Ayton & McClelland, 1997). If the Brunswikians have their way, overconfidence would be considered an artifact of the environment: If the assumptions of representative sampling are met, overconfidence disappears (Juslin, 1994). In contrast, one interpretation of the Thurstonian position might suggest that overconfidence is an artifact of response error and the resulting

regression effects (Keren, 1997).[6] The present research suggests that both of these accounts are simultaneously correct, yet misleading.

The Brunswikian account is correct in that overconfidence often is reduced or eliminated when representative sampling is used. The Thurstonian account is correct in that the reduction of random error reduces overconfidence. However, both accounts are misleading because they ask the wrong question. The Brunswikians should ask "why does representative sampling reduce overconfidence?" And the Thurstonians should ask "what psychological variables affect random error (cf. Budescu et al., 1997)?" The present research has provided answers to these questions. Representative sampling works because the questions over which participants are tested are more likely to be represented in memory. Random error is affected both by how well information is stored in memory (encoding quality) and by the amount of experience the participant has in the judgment domain (trace frequencies).

On the face of it, these two models appear to be at opposing sides of the ecological versus error debate. Which theory is "correct" and what can MDM contribute to this debate?

The most important insight provided by MDM is that there is much more common ground between the Brunswikian and the Thurstonian approaches than one might have previously believed. The same factors in the environment that the Brunswikians argued were necessary for good calibration also lead to a reduction of

[6] In addition, the analyses by Erev et al. (1994) suggest that overconfidence might be due to how data are aggregated in overconfidence analyses (cf. Yates, Lee, & Bush, 1997). In reanalyzing data from several experiments, they found that the same data set can show overconfidence if the data are

random error. Thus, contrary to previous discussions, these two accounts of overconfidence are not at odds, they are complementary. More importantly, MDM goes beyond both models by specifying the memory mechanisms responsible for both good and poor calibration. MDM goes beyond the ecological model of Gigerenzer et al. (1991) by showing how some aspects of the environment (i.e., cue structure and experience) can be instantiated in the context of a multiple-trace memory model. MDM also goes beyond the error model of Erev et al. (1994) by proposing that natural memory processes might be the source of the random error underlying judgment. Furthermore, the present research, and recent research by Wallsten et al. (1999), suggest that overconfidence is more than a data-analytic artifact of random error in judgment, and that random error results from fundamental psychological processes. To the extent that this error can be reduced through improved encoding and increased experience, overconfidence can be reduced.

Relation to Previous Findings

The present research and theory have several implications, both with respect to previous findings in decision making and with respect to debiasing overconfidence.

Hard/easy effect. A robust finding in the decision making literature is the hard/easy effect, whereby participants often are overconfident when answering difficult questions but underconfident or well calibrated when answering easy questions (Lichtenstein & Fischhoff, 1977). Previously, the hard/easy effect has been explained away by appealing to the structure of the environment or the biased selection of general

---

aggregated conditional on confidence, and underconfidence if the data are aggregated conditional on the proportion correct.

knowledge questions. Gigerenzer et al. (1991) and Juslin (1994) have argued that the hard/easy effect is a by-product of a biased selection of general knowledge questions. If the experimenter intentionally selects a difficult set of questions, participants will show overconfidence and if an easy set of questions is selected, the participant will show underconfidence. These same researchers also argue that if representative sampling is used to select the set of general knowledge questions (i.e., the questions are sampled randomly from the participants' ecological reference class), participants will be relatively well calibrated.

Studies using random sampling have had varying success in eliminating the overconfidence effect. Whereas several studies have demonstrated that overconfidence disappears with representative sampling (see Juslin, 1994), other studies have failed to find this effect (Brenner et al., 1996; Griffin & Tversky, 1992). Consequently, researchers studying overconfidence are left wondering what factors underlie the hard/easy effect.

MDM's explanation of the hard/easy effect is somewhat more explicit than previous accounts and appeals to the retrievability of items from memory. One way to interpret the hard-question / easy-question distinction is that domains from which hard questions are drawn are those in which participants have little exposure. In contrast, the domains from which easy questions are drawn are those in which participants have extensive exposure. Consequently, the answers to easy questions will be relatively more easily retrieved than answers to difficult questions. As the present research has shown, overconfidence is lower for items that are more easily retrievable (as operationalized by

encoding quality and trace frequency). Thus, one interpretation of the hard/easy effect attributes the finding to memory retrieval processes. Overconfidence will obtain when the to-be-judged items are unlikely to be represented in memory, and good calibration (and possibly underconfidence) will obtain when the majority of the items are represented in memory.[7]

One appeal of using the retrievability of items from memory as the explanation of the hard/easy effect is that it can explain why some studies using representative sampling have found overconfidence (e.g., Brenner et al., 1996; Griffin & Tversky, 1992). If too broad a reference class is chosen, whose items are beyond what is commonly experienced by the participants, overconfidence is likely to obtain. This is because few of the items will be represented in memory, which will in turn lead to a low proportion correct. If a reference class is selected that contains items commonly experienced by the participants, and therefore represented in memory, then little overconfidence would be expected.

Validity effect. A finding related to the overconfidence effect is the validity effect (Hasher, Goldstein, & Toppino, 1977). The validity effect is the tendency to rate familiar items as more valid than unfamiliar items, regardless of whether the items are true (Boehm, 1994). Factors such as the number of exposures to the stimulus (Hasher et al., 1977) and encoding (Begg, Armour, & Kerr, 1985) have been found to increase

[7] This is not to say that item difficulty is not important. Rather, item difficulty is often, if not always, confounded with whether the items are represented in memory. Perhaps one reason that representative sampling works is that it often results in a set of general knowledge items that has an equal mix of items that are represented and not represented in memory.

judgments of validity. In essence, increasing the familiarity of the stimulus increases participant's judgments of validity.

The present findings, in particular those of Experiment 3, are consistent with this prior research on the validity effect. Participant's confidence in their judgments generally increased as trace frequency increased and as encoding quality increased (cf. Begg et al. 1985; Hasher et al., 1977; Kelley & Lindsey, 1993). This result is also consistent with Dougherty et al.'s (1999) biased encoding and biased experience explanations of the validity effect.

Debiasing overconfidence. Another implication of the present research involves debiasing judgment. Previous attempts to debias overconfidence have generally been directed at the post-retrieval stage of judgment (during the retrieval stage, or during the confidence assessment stage). For example, Arkes, Christensen, Lai, and Blumer (1987) found that overconfidence could be reduced by encouraging an anchoring and adjustment strategy. Koriat, Lichtenstein, and Fischhoff (1980) were somewhat successful at reducing overconfidence by enticing participants to think of reasons why their answers might be wrong (for a review, see Arkes, 1991). In contrast to these previous attempts, the present research suggests that efforts to debias judgment should be directed at improving the initial encoding of information and/or by providing more experience in the judgment domain.

In prior research, the effect of experience on calibration was operationalized in terms of the participant's expertise in a domain, with little regard for the predictability of the task (Oskamp, 1965; Yates, McDaniel, & Brown, 1991). The present research,

and that of others using the ecological approach, suggest that the structure of the task is paramount to whether one will show overconfidence. In the context of MDM, one can expect calibration to improve with experience only in repetitive tasks that comprise essentially similar stimuli (Keren, 1991). Dynamic, and relatively unpredictable, tasks involving unique multi-dimensional stimuli (e.g., persons in clinical diagnosis; stocks in stock forecasting) do not provide the type of learning opportunity needed for calibration to improve with experience (Shanteau, 1992). Thus, in these types of tasks, experience actually leads to an increase in the frequency of *dissimilar* traces (few of the stimuli are repeated). As pointed out by Gronlund & Elam (1994), increasing the number of traces orthogonal to the target leads to an increase in variance at memory retrieval.

Obviously, improved encoding and increased experience have limited applicability in the real world. For example, how can one entice a person to engage in elaborative rehearsal strategies? Moreover, in many situations, the decision maker is confronted with only one chance to experience an event, as is the case when witnessing a crime. One way to improve calibration in these situations might be to implement strategies that enhance the retrievability of information from memory. This might be accomplished through reinstating the encoding context or by using a cognitive-reconstructive technique such as the cognitive interview (Klein, Calderwood, & MacGregor, 1989). I know of no such study that has investigated these types of techniques as debiasing methods.

Summary

The present paper has provided an account of overconfidence that is grounded in memory theory. In so doing, I have shown how a multiple-trace memory model can be used to integrate the two major theoretical accounts of overconfidence: MDM simultaneously accounts for the Brunswikian notions of ecological structure and experience and the Thurstonian notion of response variability. Although it might be tempting to conclude that overconfidence is attributable solely to memory processes, the theoretical account provided by MDM does not explain all overconfidence findings. For example, the model does not explain why there appear to be large and systematic cultural differences in overconfidence (Yates et al., 1997; Yates, Zhu, Ronis, Wang, Shinotsuka, & Toda, 1989), nor can it be readily applied to predictive judgments that do not rely on one's past memory. However, despite these shortcomings, the model and data presented here illustrate that simple memory processes go a long way towards accounting for both good and poor calibration.

References

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486 - 498.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39,* 133 - 144.

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Making, 10,* 279 - 285.

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science, 17,* 199 - 214.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes, 58,* 386 - 405.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin, 20,* 285 - 293.

Boronat, C. B., & Logan, G. D. (1997). The role of attention in automatization: Does attention operate at encoding, or retrieval, or both? *Memory & Cognition, 25,* 36 - 46.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1 - 3.

Brunswik, E. (1964). Scope and aspects of the cognitive problem. In *Contemporary Approaches to Cognition* (pp. 4 - 31). Cambridge, MA: Harvard University Press.

Budescu, D. V., Erev, I. & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making, 10,* 157 - 172.

Caccioppo, J .T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116 - 131.

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

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125,* 159 - 180.

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.

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.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650 - 669.

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

Goldstein, D. G. (1998). Inference from ignorance: The recognition heuristic. In M. A. Gernsbacher (Ed.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (pp. 407 - 411).* Mahwah, NJ: Erlbaum.

Griffin, D. W. & Varey, C. A. (1996). Towards a consensus on overconfidence. *Organizational Behavior and Human Decision Processes, 65,* 227 - 231.

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

Gronlund, S. D., & Elam, L. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355 - 1369.

Gronlund, S. D., Ohrt, D. D., Dougherty, M. R. P., Perry, J. L., & Manning, C. A. (1998). Role of memory in air traffic control. *Journal of Experimental Psychology: Applied, 4,* 263 - 280.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16,* 107 - 112.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 96,* 528 - 551.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5,* 55 - 71.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226 - 246.

Juslin, P., Olssen, 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.

Juslin, P., Olssen, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304 - 1316.

Juslin, P., Winman, A., & Persson, T. (1995). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Cognition, 54,* 99 - 130.

Kelley, C. M., & Lindsey, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32,* 1 - 24.

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

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217 - 273.

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making, 10,* 269 - 278.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19,* 462 - 472.

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

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's (eds.) *Judgment under uncertainty: Heuristics and biases,* (pp. 306 - 334). New York, NY: Cambridge.

MacDonald, A. P. (1970). Revised scale for ambiguity tolerance: Reliability and validity. *Psychological Reports, 26,* 791 - 798.

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

Mladenovic, R., & Simnett, R. (1994). Examination of contextual effects and changes in task predictability on auditor calibration. *Behavioral Accounting Research, 6,* 178 - 203.

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

Oskamp, S. (1965). Overconfidence in case study judgments. *Journal of Consulting Psychology, 29,* 261 - 265.

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.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252 - 266.

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.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273 - 286.

Tomassini, L. A., Solomon, I., Romney, M. B., & Krogstad, J. L. (1982). Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior and Human Decision Processes, 30,* 391 - 406.

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition, 28,* 277 - 296.

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

Wallsten, T. S., Bender, R. H., & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 96 - 115.

Yates, J. F. (1982). External correspondence: Decomposition of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132 - 156.

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

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

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes, 49,* 60 - 79.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes, 43,* 145 - 171.

Table 1. Frequency of Traces for Each Hypothesis-Data (Disease-Symptom) Combination in the Low and High (in parentheses) Conditions and the Objective Conditional Probabilities for the Simulations and for Experiments 1 and 2.

	H_1 (Disease 1)	H_2 (Disease 2)	$P(Disease_1 \mid Symptom)$
D_1 (Symptom 1)	9 (27)	1 (3)	.9
D_2 (Symptom 2)	8 (24)	2 (6)	.8
D_3 (Symptom 3)	7 (21)	3 (9)	.7
D_4 (Symptom 4)	6 (18)	4 (12)	.6
D_5 (Symptom 5)	4 (12)	6 (18)	.4
D_6 (Symptom 6)	3 (9)	7 (21)	.3
D_7 (Symptom 7)	2 (6)	8 (24)	.2
D_8 (Symptom 8)	1 (3)	9 (27)	.1
Total	40 (120)	40 (120)	

Table 2. Means and Standard Deviations (in parentheses) for the Probability Score and Yates' Decomposition for Experiment 1.

	Encoding			
	High n=40	Low n=39	Effect Size (d)	F (1, 78)
Mean Confidence	.708	.694	.22	0.85
	(.053)	(.078)		
Mean % Correct	.652	.578	.78	11.28**
	(.107)	(.087)		
Probability Score	.228	.261	.73	10.47**
	(.044)	(.046)		
Knowledge	.215	.236	.81	12.77**
	(.032)	(.016)		
Bias	.056	.116	.60	6.87**
	(.111)	(.090)		
Slope	.032	.018	.39	2.96*
	(.039)	(.033)		
Calibration-in-the-large	.015	.021	.25	1.26
	(.019)	(.028)		
Scatter	.011	.011	.0	0.10
	(.006)	(.006)		

* $p<.10$, ** $p < .01$

Table 3. The Means and Standard Deviations (in Parentheses) of the Probability Score and Yates' Decomposition for the High and Low Experience Groups in Experiment 2.

	Experience			
	High n = 29	Low n = 27	Effect Size (d)	F(1, 54)
Mean Confidence	.719	.717	.03	0.01
	(.062)	(.070)		
Mean % Correct	.629	.555	.76	7.95**
	(.106)	(.087)		
Probability Score	.240	.275	.70	6.62*
	(.061)	(.039)		
Knowledge	.222	.239	.81	7.93**
	(.029)	(.013)		
Bias	.090	.162	.72	6.31*
	(.121)	(.088)		
Slope	.042	.022	.42	2.36
	(.043)	(.050)		
Calibration-in-the-large	.022	.033	.33	0.45
	(.037)	(.030)		
Scatter	.013	.012	.16	0.50
	(.006)	(.006)		

* $p < .05$, ** $p < .01$

Table 4. Means and Standard Deviations (in Parentheses) of the Probability Scores and Yates's decomposition for the Good Encoding (top half) and the Poor Encoding Conditions (bottom half) and the Three Levels of Trace Frequency.

	Presentation Frequency		
Good Encoding (n = 26)	0	1	3
Mean Confidence	.740	.884	.954
	(.093)	(.054)	(.038)
Mean % Correct	.527	.820	.943
	(.086)	(.098)	(.044)
Probability Score	.314	.133	.049
	(.072)	(.066)	(.033)
Knowledge	.242	.137	.051
	(.011)	(.054)	(.036)
Bias	.213	.064	.011
	(.139)	(.088)	(.034)
Slope	.024	.130	.102
	(.042)	(.084)	(.109)
Calibration-in-the-large	.064	.011	.001
	(.064)	(.019)	(.001)
Scatter	.018	.016	.005
	(.011)	(.008)	(.004)
Poor Encoding (n=24)			
Mean Confidence	.757	.826	.868
	(.094)	(.062)	(.065)
Mean % Correct	.479	.649	.791
	(.075)	(.112)	(.120)
Probability Score	.345	.253	.163
	(.062)	(.078)	(.088)
Knowledge	.244	.215	.151
	(.006)	(.034)	(.062)
Bias	.277	.176	.077
	(.106)	(.119)	(.109)
Slope	.009	.068	.100
	(.043)	(.049)	(.071)
Calibration-in-the-large	.087	.044	.017
	(.057)	(.046)	(.030)
Scatter	.017	.019	.016
	(.010)	(.010)	(.008)

Figure Captions

Figure 1. Schematic of the conditional retrieval process in MDM.

Figure 2. Conceptual model of MDM's applicability to overconfidence.

Figure 3. MDM simulations of the effect of encoding (top panel) and experience (bottom panel) on the calibration of probability judgments. Notice that calibration is better when encoding quality is good (dashed line top panel) and when experience is high (dashed line bottom panel).

Figure 4. Effect of encoding manipulation in Experiment 1. Dashed line illustrates the good-encoding condition and solid line illustrates poor-encoding condition.

Figure 5. Effect of experience manipulation in Experiment 2. Dashed line illustrates the high-experience condition and solid line illustrates the low-experience condition.

Figure 6. Effect of encoding and experience in the general-knowledge task used in Experiment 3.

Figure 1

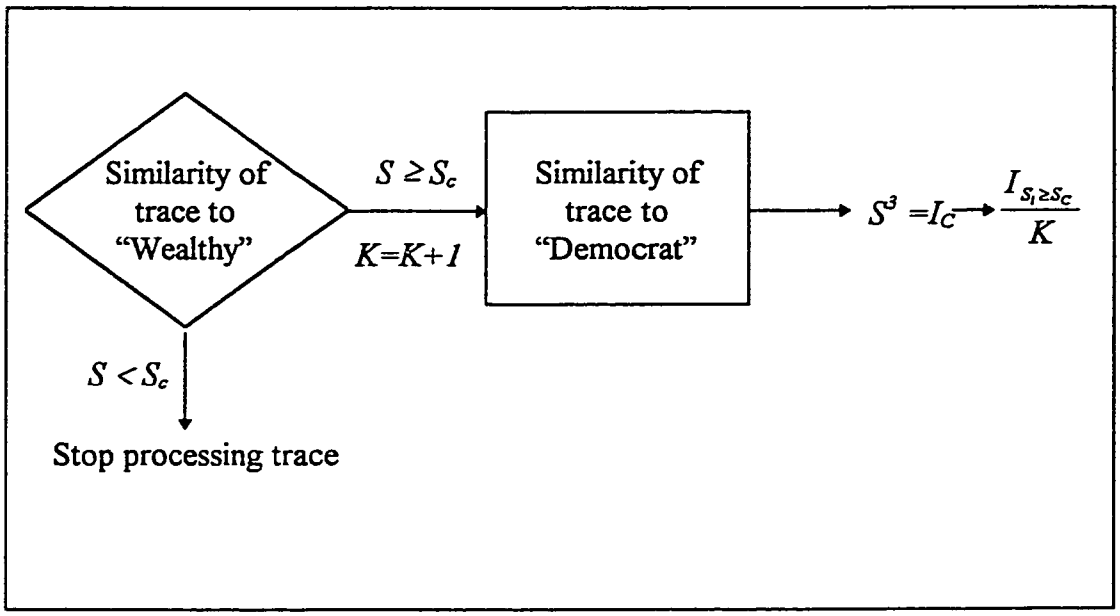

Similarity of trace to "Wealthy"
$S \geq S_c$
$K=K+1$
Similarity of trace to "Democrat"
$S^3 = I_C \rightarrow \frac{I_{S_i \geq S_C}}{K}$
$S < S_c$
Stop processing trace

Figure 2

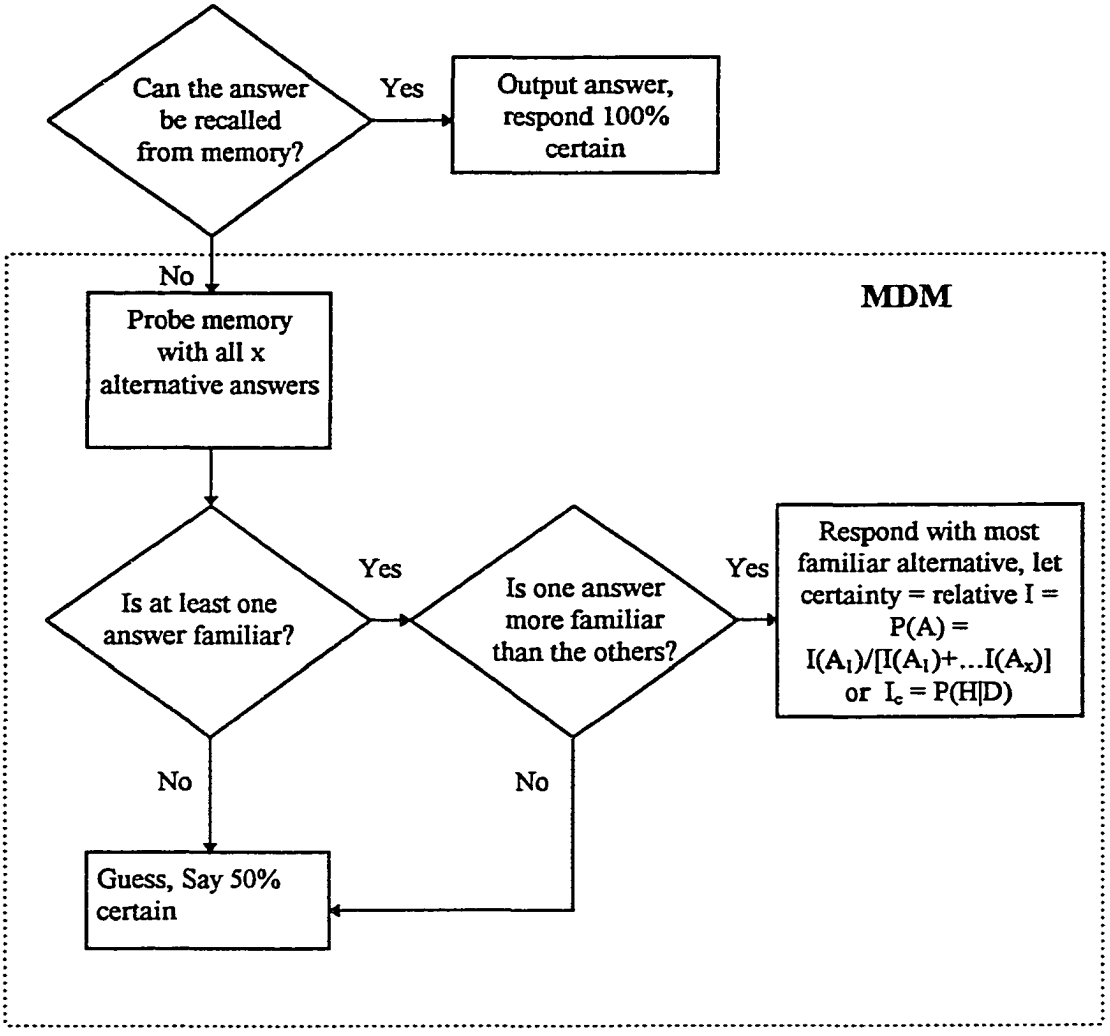
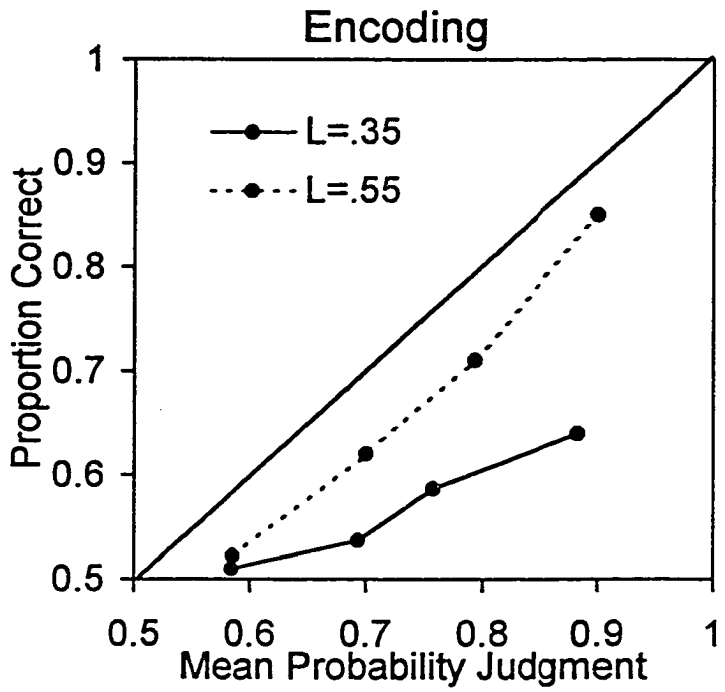

Can the answer be recalled from memory?
Yes
Output answer, respond 100% certain
No
MDM
Probe memory with all x alternative answers
Is at least one answer familiar?
Yes
Is one answer more familiar than the others?
Yes
Respond with most familiar alternative, let certainty = relative I = P(A) = I(A1)/[I(A1)+...I(Ax)] or Ic = P(H|D)
No
No
Guess, Say 50% certain

Figure 3


Encoding
L=.35
L=.55
Proportion Correct
Mean Probability Judgment
0.5
0.6
0.7
0.8
0.9
1

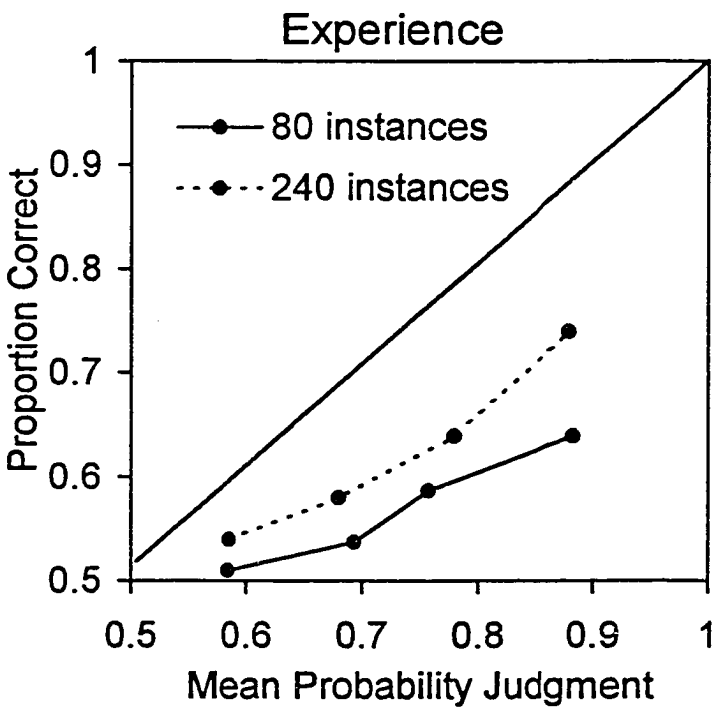

Experience
80 instances
240 instances
Proportion Correct
Mean Probability Judgment
0.5
0.6
0.7
0.8
0.9
1

Figure 4

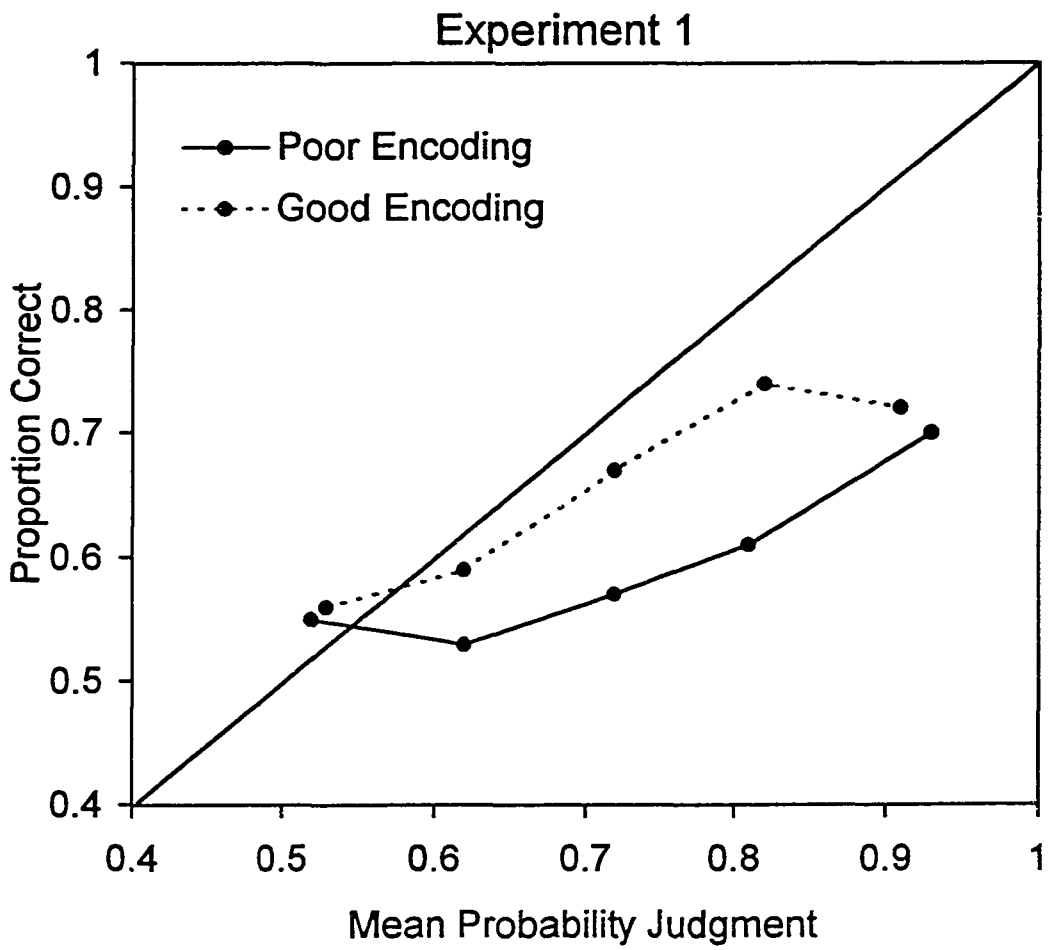

Experiment 1
Poor Encoding
Good Encoding
Proportion Correct
Mean Probability Judgment
1
0.9
0.8
0.7
0.6
0.5
0.4
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 5

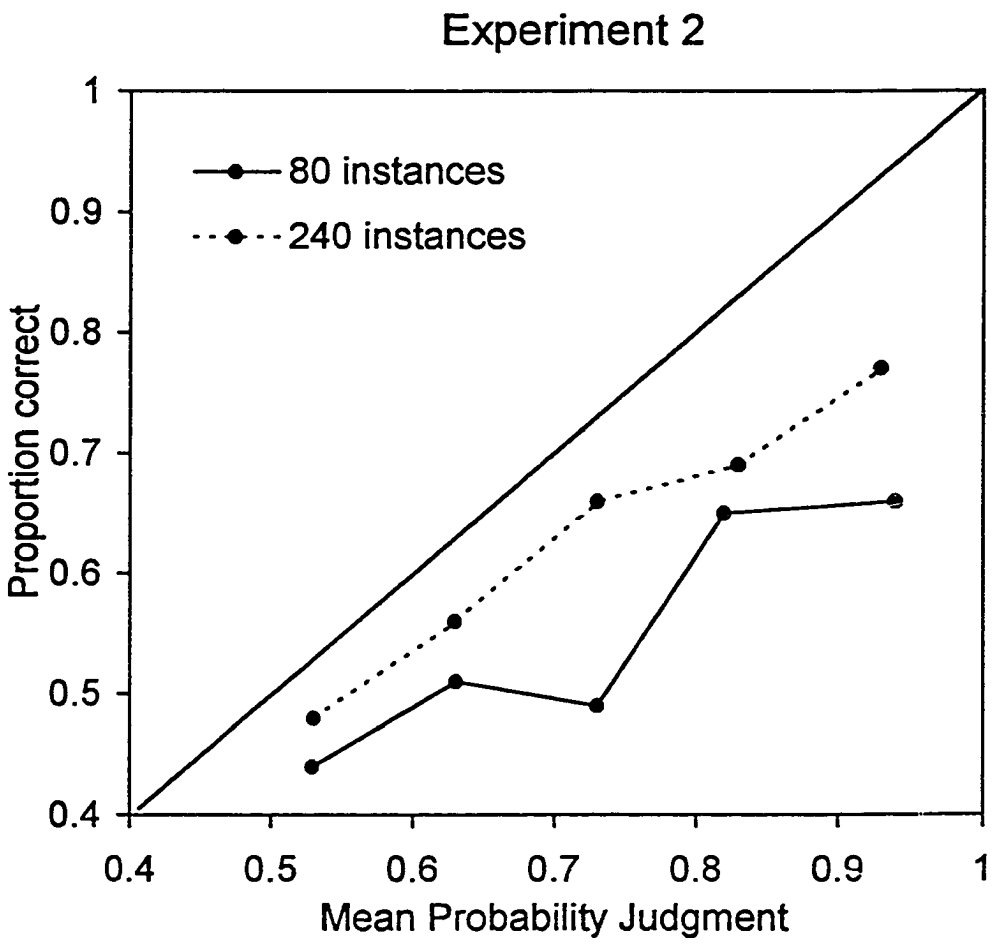

Experiment 2
80 instances
240 instances
Proportion correct
Mean Probability Judgment
1
0.9
0.8
0.7
0.6
0.5
0.4
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 6

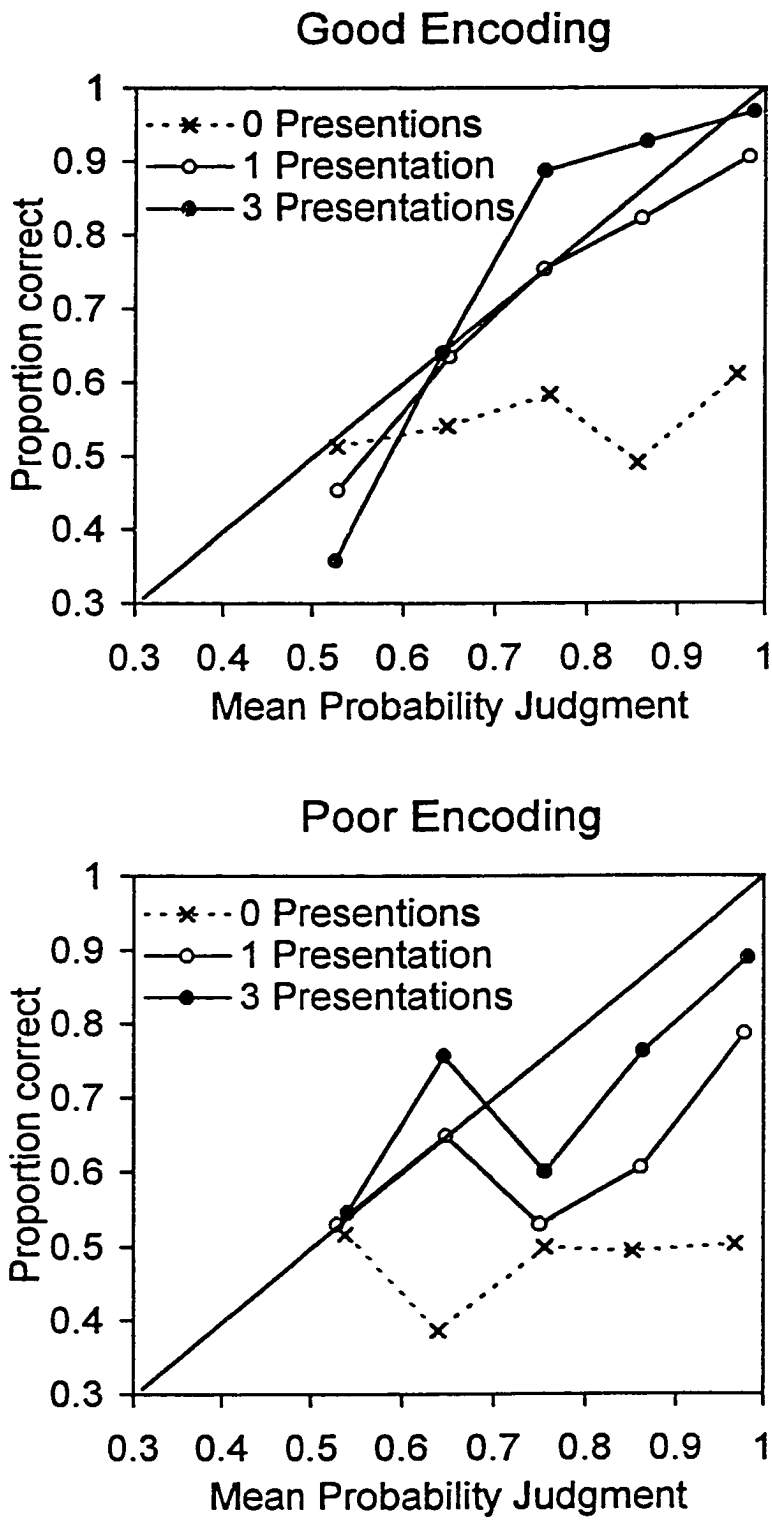
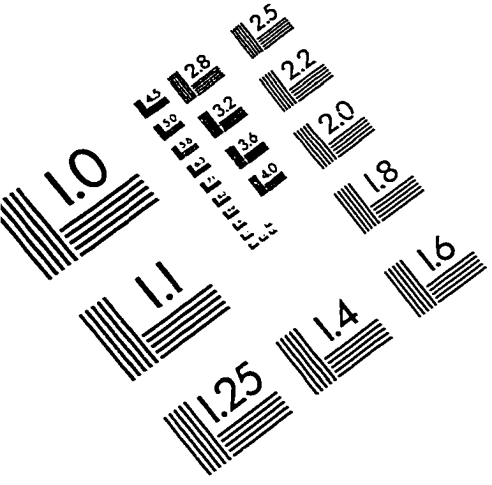
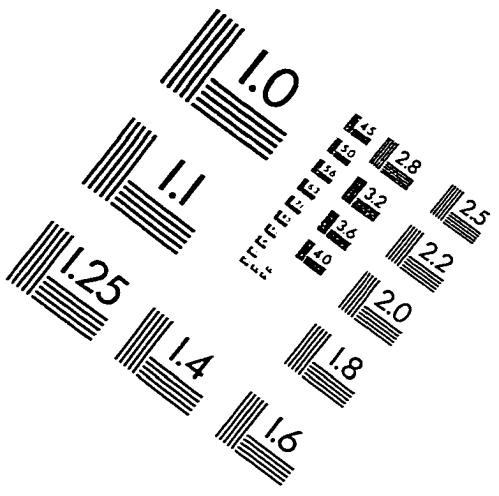
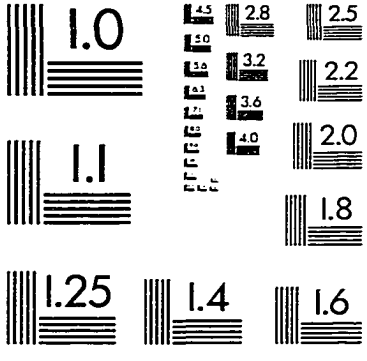
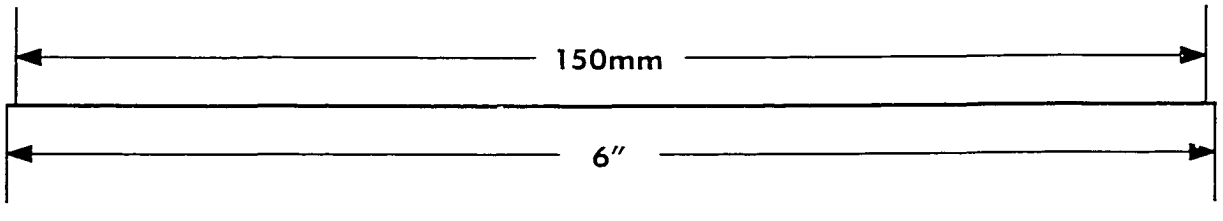

Good Encoding
0 Presentions
1 Presentation
3 Presentations
Proportion correct
Mean Probability Judgment
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Poor Encoding
0 Presentions
1 Presentation
3 Presentations
Proportion correct
Mean Probability Judgment
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

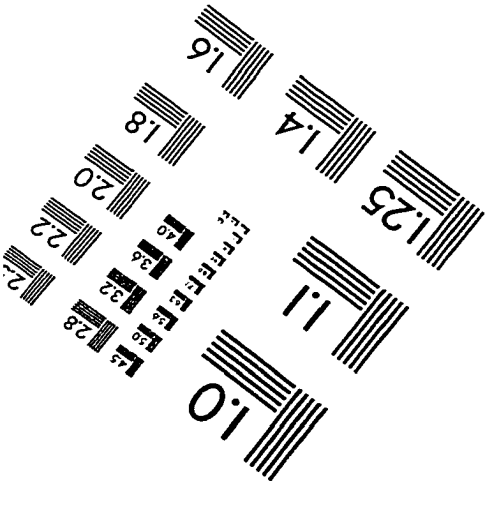
IMAGE EVALUATION TEST TARGET (QA-3)

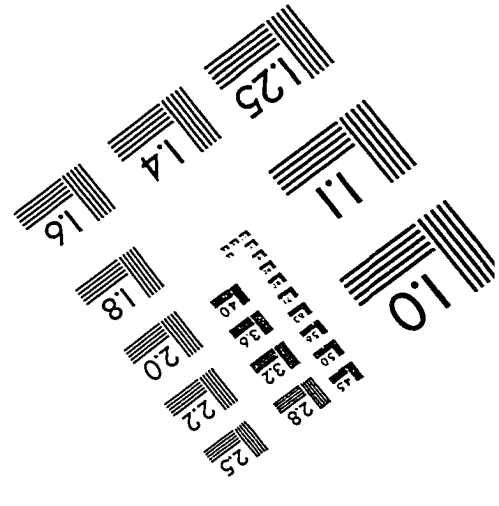












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989