Memory & Cognition
1993, 21 (2), 198-209

Do we know what we've learned from listening to the news?

SANDRA L. SCHNEIDER and SUZANNE K. LAURION
University of South Florida, Tampa, Florida

This study investigates the relationship between knowledge acquisition and an awareness of that knowledge within the context of listening to the news. Subjects listened to a recording of a radio news program consisting of regular news items as well as editorials, manipulated to be of high or low personal relevance. They then completed a surprise memory test and rated their confidence in their answers. In contrast to many studies, the results indicated a strong positive confidence-accuracy relationship. Confidence ratings were generally a better predictor of an individual's performance than were predictions based on item difficulty. Whereas subjects reported strong and accurate feelings of knowing, they apparently lacked complementary feelings of *not* knowing. The implications of these findings and others are discussed.

The average person spends some time almost every day listening to the news. But how much do we learn from listening to the news, and are we aware of what we've learned? Typically studies of knowledge acquisition and metamemory focus on classroom learning or other forms of explicit intentional learning. The present study asks some of the same types of questions, but focuses instead on a more incidental type of learning common to the routine of people's daily lives.

In particular, this study investigates learning within the context of a radio news broadcast, giving special attention to listeners' metamemory, or their awareness of what they have learned from the broadcast. Given that knowledge acquisition in both formal and informal settings requires active processing of available information, the study also examines the influence of issue involvement (or personal relevance) on memory, both for traditional informational news items and for persuasive editorial messages.

Awareness of Our Knowledge

Recent interest in people's subjective awareness of cognitive contents and processes is evident across several lines of research. These include studies of the relation between confidence judgments and accuracy, the calibration of comprehension ratings with performance, and the relation between feelings of knowing and actual memory. Most of these studies cast doubt on people's ability to accurately assess what they know. Some measures suggest that there is virtually no relationship between people's confidence in their performance and their accuracy. Other measures suggest that there is at least a weak positive relationship, but typically there is plenty of room for improvement.

Calibration of confidence judgments. Researchers in the area of judgment and decision making have largely concluded that people tend to be overconfident in their assessments of what they know (e.g., Einhorn, 1980; Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein, Fischhoff, & Phillips, 1982; Ronis & Yates, 1987). In a classic set of studies examining the confidence-accuracy relationship across a wide variety of tasks, Lichtenstein and Fischhoff (1977) showed that when knowledge and experience were lacking, there was virtually no relationship between confidence and accuracy. In other more familiar tasks, they found that most subjects were prone to overconfidence, especially subjects who performed poorly. The strength of the confidence-accuracy relationship varied as a function of item difficulty, with subjects who knew more tending to be more confident of their responses. Lichtenstein and Fischhoff concluded that the confidence-accuracy relationship is likely to be best calibrated at about 80% accuracy levels but that, in general, subjects seem relatively insensitive to how much they really know.

Attempts to improve confidence-accuracy calibration have met with only limited success. Lichtenstein and Fischhoff (1977, Experiment 2; 1980) found that training subjects in an unfamiliar task not only improved performance accuracy but also led to improved confidence-accuracy calibration. Koriat, Lichtenstein, and Fischhoff (1980) successfully reduced overconfidence by having subjects consider in what ways their answers might be wrong. However, Sniezek, Paese, and Switzer (1990) did not find a comparable reduction in overconfidence when they asked their subjects to judge the probability of being wrong rather than judging the probability of being correct.

Using a variant of the feedback approach, Arkes, Christensen, Lai, and Blumer (1987) provided subjects with apparently easy, but misleading, items and then gave

Many thanks to Jim Jenkins for his comments and encouragement on an earlier draft of this work. We would also like to thank two anonymous reviewers for their helpful comments. Portions of these data were presented at the 32nd Annual Meeting of the Psychonomic Society, San Francisco, in November 1991. Correspondence regarding this article should be addressed to Sandra L. Schneider, University of South Florida, Department of Psychology, BEH 339, Tampa, FL 33620-8200.

them feedback on their accuracy. The primary effect of learning about these "trick" questions on later judgments was to introduce underconfidence. Yet the presence of feedback, in both this condition and another, led to slight improvements in calibration. In a second experiment, calibration was mildly improved by informing subjects that they would be asked to justify their answers to a group. Tetlock and Kim (1987) have also found that alerting subjects to the need to justify their responses improves confidence-accuracy calibration, provided that subjects are warned of their accountability prior to viewing the material to be judged.

Not all studies show the tendency toward overconfidence in judgment. Lichtenstein and Fischhoff (1977, Figure 10) found that graduate students were highly calibrated in their confidence judgments when responding both to psychology items and to general knowledge items. Investigations of professionals, such as meteorologists and auditors, who regularly estimate probabilities of occurrences, have generally produced strong positive relationships between job-related probability estimates and the actual frequency of occurrence of predicted events (Murphy & Winkler, 1977, 1984; Tomassini, Solomon, Romney, & Krogstad, 1982). Even among bridge players, Keren (1987) has documented that experts are well-calibrated in their probability judgments, whereas amateurs demonstrate the typical overconfidence effect.

Calibration of comprehension. Consistent with the general findings from judgment research, investigations of metacognitive processes in comprehension tasks have produced what is commonly referred to as an "illusion of knowing." Essentially, these studies suggest that there is only a weak relationship between our sense of what we comprehend from expository text and how well we actually perform on tests of comprehension.

In a series of studies, Glenberg, Epstein, and their colleagues asked subjects to read informative text materials and to rate their confidence that they understood the texts. In the initial studies, subjects were warned that the texts might contain contradictions and were explicitly told to look for them. Frequently, subjects failed to find the contradictions, yet maintained high levels of confidence that they had understood the passage. This failure to detect contradictions was evident across a wide variety of conditions despite the fact that subjects were aware that contradictions would probably be present (Epstein, Glenberg, & Bradley, 1984; Glenberg, Wilkinson, & Epstein, 1982).

In later studies, Glenberg and Epstein measured comprehension using a single confidence rating that was made after reading a particular (noncontradictory) text, but before being exposed to the true-false question of interest. Using correlations between confidence ratings and accuracy as a measure of the calibration of comprehension, several of their studies documented that there was virtually no relationship between subjects' beliefs about their comprehension and their subsequent ability to correctly answer the comprehension questions (Glenberg & Epstein, 1985, 1987; Glenberg, Sanocki, Epstein, & Morris, 1987; Morris, 1990). The authors conclude that poor calibration of comprehension occurs because subjects seem to assess their familiarity with the topic domain rather than assessing their knowledge of facts encountered in the text.

Using a similar technique, Maki and Berry (1984) have also found only weak (and frequently nonsignificant) correlations between subjects' confidence in their comprehension of psychology text material and later test performance. However, in follow-up research, Maki and her colleagues have found some evidence that subjects can predict performance to a certain extent, especially when text processing is enhanced or ratings are taken after the test has been completed (Maki, Foley, Kajer, Thompson, & Willert, 1990).

Similarly, Glenberg and Epstein (1985, Experiment 3) found that the confidence-accuracy relationship could be improved by asking subjects to judge their confidence at the same time they answer the comprehension questions. They referred to this measure as the "calibration of performance" because the subjects made their confidence ratings only after being exposed to, and answering, a particular question. Even here, however, although the correlation was statistically different from zero (r_{pb} = .23), the confidence-accuracy relationship was still relatively weak and far from perfect. In a similar follow-up experiment, Glenberg and Epstein (1987) found moderately good calibration of performance (with average Goodman-Kruskal gamma correlations of .42 and .36 in two different conditions).

Recently, Weaver (1990) has suggested that calibration may be substantially higher than is generally suggested by Glenberg, Epstein, and their colleagues when more reliable measures of calibration are used. To make his case, Weaver showed that tests of calibration using only a single question per topic systematically underestimate the strength of the confidence-accuracy relationship.

Using multiple-item tests of various types, Pressley and Ghatala (1988) have demonstrated that, in some contexts, subjects can predict their performance at well above chance levels. However, they also found that subjects maintained a troublesome tendency toward overconfidence in incorrect answers, especially in comprehension tests, and that performance predictions tended to become less accurate as item difficulty increased.

Feeling of knowing. In contrast to most of the previously cited studies of calibration, investigations of subjective feelings of knowing are frequently presented as evidence that people have substantial insight into their own cognitive processes. A feeling-of-knowing judgment consists of an assessment by an individual of the likelihood that he or she will be able to recognize a given fact or association when he/she is unable to recall it.

In an early study of the feeling of knowing, Hart (1967) presented subjects with general knowledge questions, such as "Which planet is the largest in our solar system?" When subjects could not produce the answer to a question, they were asked to report whether they experienced a feeling of knowing (i.e., a feeling that they knew the

answer but simply could not recall it at the moment). After the initial test, they were given a second version of the same test, but this time in a four-alternative multiple-choice format in which only recognition of the correct answer was required.

Hart (1967) found that subjects were reliably more likely to recognize the correct answer when they reported a feeling of knowing (76% and 66% in two experiments) than when they reported no feeling of knowing (57% and 62%). Although this suggests some metamemory ability, the differences in recognition performance were not large. Moreover, subjects were less accurate than might be expected when they reported a feeling of knowing and more accurate than might be expected when they reported no feeling of knowing. Nelson, Leonesio, Shimamura, Landwehr, and Narens (1982) and Schachter (1983) have also found relatively weak but reliable feeling-of-knowing effects in paired associate list-learning tasks.

Nelson, Leonesio, Landwehr, and Narens (1986) showed that a given individual's feeling-of-knowing judgments were better predictors of that person's recognition performance than were pooled feeling-of-knowing judgments from a larger group of subjects. Nevertheless, they also showed that the individual's judgments were generally not as good at predicting later performance as a normative index of item difficulty. This suggests that although subjects have some sense of what they know, it does not seem that they have substantial idiosyncratic access to the contents of their knowledge.

In a less standard test of metamemory, Vesonder and Voss (1985) found only a weak relationship between confidence and recall when subjects rated their certainty that they would be able to later recall the target word from the word pair currently being studied. However, when Lovelace (1984) manipulated whether study occurred in a distributed or massed fashion, he found that the confidence-accuracy relation for paired associate recall improved substantially with repeated, shorter study presentations in contrast to the single, longer study presentations. Recently, Nelson and Dunlosky (1991) have also reported strong relationships between confidence and later paired associate recall, but only when the confidence judgment is slightly delayed rather than immediate.

Finally, in a test of autobiographical metamemory, Barclay and Wellman (1986) found that subjects were originally fairly well calibrated in their confidence about memory for recent events in their own daily lives. However, over time, while confidence remained high, recognition accuracy declined steadily.

Gaining Knowledge From the News

On the whole, then, the case for metacognitive skills with regard to access of knowledge seems mediocre at best. Studies of people's awareness of their knowledge, however, have focused primarily on general knowledge, classroom material, or laboratory list-learning exercises. To date, investigations have rarely touched on knowledge gained within typical day-to-day experiences. The few exceptions include investigations of job-related predictions (e.g., Murphy & Winkler, 1977) and autobiographical memory (e.g., Barclay & Wellman, 1986). Interestingly, there is at least some suggestion that in these contexts the confidence-accuracy relationship may be stronger than in others.

The present investigation focuses on listening to the news: a routine daily activity that nonetheless occurs primarily for the purpose of acquiring information. Unlike typical instances of intentional learning, the motivation to learn when listening to the news is not dictated by perceived expectations of an instructor or an experimenter, but instead is the product of what the listener wants or chooses to process. The type of processing itself is less formally constrained and less prone to conscious strategic attempts to successfully encode the material. In addition, the information available for processing in a news broadcast is generally more intrinsically meaningful and more concrete, focusing on contemporary and forthcoming events rather than on general principles, facts, or simple word associations.

This contrast suggests that the outcome of listening to the news may differ substantially from the typical findings previously summarized. While it is not entirely clear whether subjects will learn more or less in such an informal learning environment, we hypothesized that they would have relatively accurate insight into what they had learned, primarily because the learning would be a product of self-initiated processing.

In the present study, journalism students were asked to listen to a tape of a radio news broadcast, ostensibly for the purpose of evaluating a potential new radio station. After hearing the broadcast, they were given a surprise memory test in a four-alternative multiple-choice format. After answering each question, the subjects were asked to use a 7-point scale to indicate their confidence that they had answered the question correctly.

This method of assessing recognition memory and the confidence-accuracy relationship has two primary advantages over other methods. First, recall measures do not accurately reflect all of the knowledge that a person has regarding a particular topic.[1] This fact is the cornerstone of feeling-of-knowing research. Because we wanted to measure knowledge in a way that was reliable and easy to compare across subjects, we opted for recognition performance. Second, given the generally poor relationships between accuracy and confidence that are all too common in the literature, we wanted to select a task where there would be some precedent for expecting at least some relationship between confidence and accuracy in other contexts. Although alternative measures of memory would also be informative, the view taken in the present study is that a news message can be considered effective if the audience hearing it has encoded and stored the informa-

tion to the extent that they could recognize facts as correct or incorrect in a subsequent interpersonal or mass media communication.

Involvement and Persuasion in Learning

Beyond the measure of memory, there is also the issue of identifying factors within the media context that may contribute to successful memory and realistic metamemory. One of the most reliable influences on message processing occurs as a result of the perceived relevance of the message to the audience. When members of the audience feel that a message may be relevant to events in their own lives, they are more apt to experience a sense of issue involvement (Apsler & Sears, 1968). Issue involvement, in turn, has been associated with more elaborate and active processing of messages.

Petty and Cacioppo (1986), in their elaboration likelihood model of persuasion, suggest that the perception of personal relevance or issue involvement increases the motivation to carefully process a persuasive message:

> As the personal consequences of an advocacy increase, it becomes more important for people to form a veridical opinion because the consequences of being incorrect are greater. Because of these greater personal consequences, people should be more motivated to engage in the cognitive work necessary to evaluate the true merits of the proposal. (p. 82)

In their model, high-involvement messages are more likely to be processed through elaboration and the activation of cognitive responses. On the other hand, little attention to content is expected for low-involvement messages.

Although the model is specifically focused on the role of elaboration in effecting attitude change, the elaboration hypothesis also has implications for memory. Indeed, the earliest studies of elaboration were concerned with the influence of semantic processing on memory and, virtually without exception, demonstrated strong and highly reliable memory advantages for information that was processed through meaningful associations rather than through more superficial means, such as rote rehearsal (Bower & Clark, 1969; Bower & Winzenz, 1970; Craik & Lockhart, 1972; Craik & Tulving, 1975; Hyde & Jenkins, 1973).

In studies of the effects of issue involvement on message processing, Petty & Cacioppo (1979a, 1979b, 1984; Petty, Cacioppo, & Goldman, 1981; Petty, Cacioppo, & Schumann, 1983) have demonstrated that subjects generate more thoughts about the arguments presented in high-involvement conditions and are more sensitive to whether those arguments are strong or weak. Frequently, argument recall is also superior in the high-involvement conditions; however, the differences are not always reliable.

In the few studies of persuasion that explore memory differences as a function of involvement, better memory is usually found for high-involvement messages. In their study of the effects of issue involvement on persuasive health-related message processing, Maheswaran and Meyers-Levy (1990) found that high-involvement subjects produced more message-related informational thoughts and fewer simple evaluative thoughts than did low-involvement subjects during a free-recall task. Within an advertising context, Andrews and Shimp (1990) found that high-involvement subjects remembered significantly more of the persuasive arguments and were more likely to remember the name of the product.

Because the purpose of most persuasion studies does not include a rigorous test of memory, the evidence for better memory of high-involvement messages is relatively weak. When memory is measured, it is usually in the form of free recall, which, as mentioned earlier, is a relatively insensitive indicator of the information that is present in memory.

In our study, we specifically set out to measure differences in recognition memory for news messages as a function of involvement. Not surprisingly, we hypothesized that subjects would remember more of the details of high-involvement messages. We also hypothesized that subjects would be more confident of what they had learned in the high-involvement condition. If subjects spend more time elaborating high-involvement messages (even if they develop new associations rather than focusing on the original facts), they should be more apt to report the experience of having knowledge about the message. This seems particularly likely if, as suggested by Glenberg et al. (1987), subjects' confidence judgments are more likely to reflect general familiarity with an issue rather than availability of relevant facts.

Following Apsler and Sears (1968) and Petty and Cacioppo (e.g., Petty & Cacioppo, 1979a), we presented all subjects with the same basic message content, but high-involvement subjects were led to believe that the event described in the message would affect them personally, whereas low-involvement subjects were led to believe that the depicted event would not affect them. This manipulation avoids the interpretive problems encountered when using different messages across conditions, such as potential differences in subjects' familiarity with various issues (Petty & Cacioppo, 1986, p. 83).

We also manipulated the format of the news messages. Previously, studies of the influence of involvement have routinely employed messages specifically designed to be persuasive. On the other hand, the type of messages that we were interested in studying were the more typical informational and relatively objective news messages. Because we had reason to believe that involvement would influence memory for persuasive editorial messages, we also hypothesized that it should have a similar impact on memory for strictly informational messages. In addition, we were curious to see if confidence about memory would differ as a function of whether an item was presented persuasively as an editorial or informationally as a more typical news item.

METHOD

Subjects

Seventy-two undergraduate students from an introductory journalism class at the University of Wisconsin-Madison volunteered

Table 1
Summary of Message Contents Used in Each Broadcast

Topic Code	Message Content
Phone	9-1-1 emergency phone system to be installed in either subjects' city or a distant city.
Break	Proposal to eliminate spring break from the school calendar effective either immediately or after the subjects graduated.
Paper	Proposal to institute a mandatory term paper policy effective immediately or after subjects graduated.
Skyway	Skyways to be constructed on either subjects' campus or a distant campus.
House	Halfway house to be built on subjects' campus or on the other side of town.
Grounds	Custodial/groundskeeper budget cuts at either subjects' campus or state office buildings.
Health	Increased student fees for health care at either subjects' college or distant college.
Texts	Petition drive to limit the number of textbooks a professor can require a student to buy at either subjects' campus or a distant campus.

Note—The choice of location or time was determined by whether the message was presented in the high- or the low-involvement condition, respectively.

to participate in the study for course credit. They participated in groups of 1 to 4 over a period of 1 week. The subjects were randomly assigned to one of eight tape conditions.

Stimuli

An effort was made in this experiment to create an environment and task that were consistent with the real world of radio listening. The eight audio tapes that served as stimuli were professionally produced to be representative of real radio programming. The experiment's cover story created an objective of broadcast evaluation, which is similar to the objective of the average radio listener who selects a radio station on the basis of judgments regarding the quality of programming.

Two experts (one of whom is the second author) with several years of combined experience in editing and directing radio news as well as teaching broadcast journalism evaluated each tape in an effort to ensure (1) that messages were representative of typical news items in quality and duration and (2) that each tape and each message was matched for intensity and rate of delivery as well as voice quality.

The messages within the tapes were constructed to be of relatively high interest to the listening audience of college freshmen. All messages were controversial, meaningful, and believable, yet unfamiliar to the subject audience. A summary of the message topics is presented in Table 1.

Design

A 2×2×8 (level of involvement × message format × tape) mixed repeated measures design was employed. The two primary independent variables, level of involvement and message format, were manipulated by creating different versions of each message. Each subject heard eight different topics, with two messages from each of four conditions: high-involvement editorial, low-involvement editorial, high-involvement informational, and low-involvement informational. Each version of a message was about the same length, was composed of nearly all the same words, and was delivered in the same manner by the same source.

As a between-subject manipulation, eight different tapes of the broadcast were created. This control manipulation was intended to balance potential topic and order effects. The order of topics within each tape was determined using Wagenaar's (1969) method for constructing N×N "digram-balanced" Latin squares. The resulting topic sequence, as well as the order of conditions, is presented in Table 2.

Level of involvement. Each message was varied to achieve relatively *high* or *low* involvement on the part of subjects. Messages in both the high- and the low-involvement conditions represented the same communication; however, in high-involvement stories, the subjects were led to believe that the advocated change would affect them, whereas in low-involvement stories, the subjects were led to believe that the change would have no personally relevant effects. Borrowing the technique used by Petty and Cacioppo (1979a), high-involvement messages were described as occurring in the near future at the subjects' own university (University of Wisconsin), whereas low-involvement messages were described as occurring either in the distant future or at a different location (e.g., University of Minnesota).

The involvement manipulation was checked in a pilot investigation. Fifteen subjects heard one of the tapes and were asked to rate each of the eight messages on a Likert scale reflecting their per-

Table 2
Latin Square Indicating the Order of Message Topics and Conditions for Each of the Eight Radio Broadcast Tapes

	Topic and Condition Order							
Tape	1st	2nd	3rd	4th	5th	6th	7th	8th
1	Grounds	Health	Break	Paper	Phone	Texts	Skyway	House
	HI/Ed	HI/Inf	LI/Ed	LI/Inf	HI/Ed	HI/Inf	LI/Ed	LI/Inf
2	Health	Paper	Grounds	Texts	Break	House	Phone	Skyway
	LI/Inf	LI/Ed	HI/Inf	HI/Ed	LI/Inf	LI/Ed	HI/Inf	HI/Ed
3	Break	Grounds	Phone	Health	Skyway	Paper	House	Texts
	HI/Ed	HI/Inf	LI/Ed	LI/Inf	HI/Ed	HI/Inf	LI/Ed	LI/Inf
4	Paper	Texts	Health	House	Grounds	Skyway	Break	Phone
	LI/Inf	LI/Ed	HI/Inf	HI/Ed	LI/Inf	LI/Ed	HI/Inf	HI/Ed
5	Phone	Break	Skyway	Grounds	House	Health	Texts	Paper
	LI/Inf	LI/Ed	HI/Inf	HI/Ed	LI/Inf	LI/Ed	HI/Inf	HI/Ed
6	Texts	House	Paper	Skyway	Health	Phone	Grounds	Break
	HI/Ed	HI/Inf	LI/Ed	LI/Inf	HI/Ed	HI/Inf	LI/Ed	LI/Inf
7	Skyway	Phone	House	Break	Texts	Grounds	Paper	Health
	LI/Inf	LI/Ed	HI/Inf	HI/Ed	LI/Inf	LI/Ed	HI/Inf	HI/Ed
8	House	Skyway	Texts	Phone	Paper	Break	Health	Grounds
	HI/Ed	HI/Inf	LI/Ed	LI/Inf	HI/Ed	HI/Inf	LI/Ed	LI/Inf

Note—See Table 1 for key to topic codes. HI = high involvement, LI = low involvement, Ed = editorial format, Inf = informational format.

ceived level of interest or involvement. Six of the eight topics successfully reflected a difference between high and low involvement and were retained. The other two topics (one about the building of a new parking ramp and one about using street barricades to nab drunken drivers) could not be successfully manipulated for involvement and were replaced by messages about placing a limit on the cost of textbooks and charging students a fee for student health. Although it was not possible to pilot the latter two messages prior to the experiment, later tests of the involvement manipulation using virtually identical messages confirmed the expected difference between high and low involvement for these two topics.

Message format. Messages were presented in either an *editorial* or an *informational* format. In the editorial format, facts and arguments were presented, along with an attempt to influence the opinion of the listener on the issue. In the informational format, facts and arguments were also presented, but with no direct attempt to influence the opinion of the listener on the issue. Although the same information was presented regardless of format, arguments in editorials were attributed to a collective "we," whereas arguments in news stories were attributed to "opponents" and "supporters."

A manipulation check was conducted prior to the experiment to confirm that listeners noticed the difference in format. Eight pilot subjects listened to a randomly selected tape and afterwards were asked to label each of the eight messages as either an "editorial" or a "news story." The subjects correctly identified an average of 6.4, or 77%, of the eight message formats.

Measuring knowledge acquisition. A set of 40 multiple-choice questions was used to assess the amount of knowledge acquired from the broadcast. The items focused on important concepts and facts central to the message. There was one correct answer to each question: the exact passage from the message. Three plausible alternatives were constructed for each question. There were five questions for each topic. The questions were rearranged for each tape condition such that items about each topic were presented in the order that the topics were heard.

Development of the questions included two pilot manipulation checks. Eight subjects participated in one pilot study to identify multiple-choice questions that were either too easy or too hard. A tape was randomly chosen and played for the subjects under experimental conditions. A 56-item multiple-choice test was then administered. Questions that were answered correctly by more than 6 of the 8 subjects were either dropped or rewritten. Questions that were answered incorrectly by at least 6 subjects were also dropped or rewritten. A set of 40 questions, 5 for each topic, was retained.

At this point, a separate set of 5 subjects participated in a pilot study to determine whether the multiple-choice questions were completely text dependent. They were asked to answer the 40 questions without hearing or knowing anything about the messages. In no case were any of the questions answered correctly by more than 2 of these subjects. However, interviews with these subjects led to minor adjustments in 3 of the 40 questions.

Measuring metamemory. As subjects answered each multiple-choice item, they also rated their confidence that their answer was correct. This was done using a 7-point Likert scale, ranging from *not at all sure* (1) to *completely sure* (7).

Other measures. At the end of the experiment, the subjects were prompted with the topic of each message and asked to recall whether the message was an editorial or a news story. For the messages they perceived to be editorials, the subjects indicated whether they agreed or disagreed with the point of view expressed. Finally, the subjects were asked to use a 7-point Likert scale to indicate how likely they would be to listen to the radio station depicted. This final measure was included primarily to maintain consistency with the cover story presented to the subjects, suggesting that the purpose of the study was to evaluate the radio station.

Procedure

The experiment was conducted in a small, bare office with the window blinds closed. The subjects sat in chairs around a table. One of eight tapes was played over a portable audio tape playback machine.

The subjects were told, via a prerecorded message, that the FCC had given approval for a new radio station in their city and the owners were considering several different formats. They were told to evaluate the news/talk format as it would sound if it were on the air today. The subjects were also told that although they were not in their home or car, they were to listen to the tape as they normally listen to radio. This introduction was delivered by a professional male broadcaster using a fictitious name. The "sample tape" of eight news messages was then played (including about 3 min of distractor material—commercials and sports report—at the end). At the conclusion, the subjects were told (again via prerecorded message) that the demonstration tape was finished and that they would now be given a questionnaire to fill out. The total running time on the tape was 14 min.

The multiple-choice test was prefaced with a written statement that a chief concern of the programmers is that their news programming be memorable and subjects should therefore do the best they can on this memory test. The test, and the follow-up message and station evaluations, took approximately 10 min to complete.

RESULTS

Two sets of analyses are presented. The first evaluates the roles of level of involvement and message format on knowledge acquisition from a radio news broadcast. As part of this analysis, the potential effects of several control variables are also examined, including characteristics of the different tapes as well as the order and content of particular messages. The role of format recognition and editorial agreement or disagreement in memory accuracy is also considered.

The second analysis focuses on the extent to which subjects' confidence in their performance accurately reflects their level of knowledge acquisition. The analysis includes an assessment of the overall relationship of confidence to accuracy as well as consideration of potential differences in calibration due to the influence of message format, level of involvement, and item difficulty.

Effects on Knowledge Acquisition

Involvement and message format. The data for these analyses were the mean number of multiple-choice questions answered correctly, given five questions for each of the eight messages. For the primary 2×2 analysis of variance (ANOVA), the data were pooled over individual messages into four conditions: high-involvement editorial, high-involvement informative, low-involvement editorial, and low-involvement informative. Because each condition included two messages, the maximum number of correct answers within any condition was 10.

There was a significant main effect of level of involvement [$F(1,71) = 8.25$, $MS_e = 2.00$, $p < .006$]. As expected, the mean number of correct answers was greater for high-involvement messages (7.22) than for low-

involvement messages (6.74). There was not, however, any effect of message format ($F < 1$, $MS_e = 0.78$), nor was there a message format × involvement interaction ($F < 1$, $MS_e = 1.55$). This suggests that knowledge acquisition will generally be greater for high-involvement messages than for low-involvement messages, regardless of the editorial or informative format of the message.

Control variables: Order, tape, and repetition. Three subsidiary analyses were conducted to determine whether the chronological order of messages, the repetition of format/involvement conditions, or the idiosyncracies of particular tapes might have influenced the findings in the primary analysis. No significant difference for any of these variables was found, with the exception of a tape × repetition interaction [$F(7,64) = 2.441$, $MS_e = 5.02$, $p < .03$]. Simple effects analysis revealed that there were significant differences in memory among the tapes in the second half of the tapes [$F(7,64) = 2.34$, $MS_e = 9.79$, $p < .04$], but not in the first ($F < 1$, $MS_e = 5.33$), and that these differences were isolated to Tape 4 [$F(1,64) = 4.43$, $MS_e = 5.02$, $p < .04$] and Tape 8 [$F(1,64) = 9.96$, $MS_e = 5.02$, $p < .003$], in which accuracy was significantly lower in the second half of the presentation than in the first.

Although an examination of the two tapes revealed no obvious reasons for the differences in accuracy between the first and second halves of each tape, the data from these two tapes were separated from the data for the remaining six tapes. A reanalysis of the six tapes that were free of repetition effects again revealed that only level of involvement significantly affected knowledge acquisition [$F(1,53) = 7.60$, $MS_e = 1.71$, $p < .009$]. A separate analysis of Tapes 4 and 8 revealed that high-involvement messages were better remembered than were low-involvement messages only for Tape 8 [$F(1,16) = 8.98$, $MS_e = 1.05$, $p < .01$], and that level of involvement had no effect on accuracy within Tape 4 [$F(1,16) = 1.33$, $MS_e = 1.05$, $p > .26$]. There were no other significant effects or interactions.

Hence, for seven of the eight tapes, there was a significant effect of involvement and there was no effect of message format for any of the tapes. The effect of repetition in two of the tapes and the lack of an effect of involvement in one of the tapes might well have been due to sampling error given that only 9 subjects listened to each of the different tapes.

Format recognition, agreement, and disagreement. Before concluding that message format has no effect on memory, it is important to be able to demonstrate that the subjects were sensitive to the format manipulation. As described previously, the subjects were asked at the end of the experiment to identify which messages were editorials and which were news stories. For the items identified as editorials, the subjects indicated whether they agreed or disagreed with the position expressed.

The format of the messages was correctly identified an average of 74.8% of the time. Although average percentages ranged from 66.7% to 84.7%, none of the eight message conditions (high involvement vs. low involvement × editorial format vs. informative format × first presentation vs. second presentation) was significantly different from the overall mean according to tests of the binomial z. These results suggest that the subjects showed evidence of recognizing and remembering the format of most messages across all of the conditions.

A *post hoc* analysis was conducted to see whether recognition accuracy might have been affected by an awareness of the format of the message or by the level of agreement with the message if it was an editorial. To do this, the eight recognition test scores for each subject were categorized into one of five groups: editorial agreed, editorial disagreed, editorial misidentified, informational identified, and informational misidentified. This five-group data set was then subjected to a series of pairwise t tests. The results revealed little evidence of differences in performance, except that the subjects showed a marginal but systematic tendency toward better memory for editorials with which they disagreed than for any of the other messages.

Confidence and Accuracy

The second set of analyses focuses on the relationship between the subjects' reported confidence in their responses and the accuracy of those responses. First, an analysis is presented in which the data have been organized to evaluate and compare levels of accuracy and confidence, again including the involvement and format variables. Second, the relationship of confidence ratings to accuracy levels is assessed, with special attention to item difficulty and performance predictability.

Analysis with involvement and format. The confidence-accuracy relationship was assessed using the Goodman-Kruskal (Goodman & Kruskal, 1954) gamma correlation. Gamma (G) indicates the probability that the ordering along one type of measure (e.g., confidence) will correctly predict the ordering on another type of measure (e.g., performance accuracy). In the current context, gamma provides a measure of relative confidence-accuracy calibration by indicating the extent to which higher levels of confidence are differentially associated with higher levels of accuracy in recognition performance. Nelson (1984) provides a convincing case that this statistic is by far the most informative and appropriate measure of the confidence-accuracy relationship.

Separate gammas were computed for each subject first for the high- and low-involvement conditions (mean $\gamma = +.76$ and $+.73$, respectively) and then for the editorial and informational format conditions (mean $\gamma = +.75$ and $+.76$, respectively). In all cases, the gammas indicated surprisingly strong positive confidence-accuracy relationships. The results of two t test comparisons indicated that this relationship did not differ significantly either as a function of involvement [$t(71) = 0.70$, $SD = 0.34$, $p > .48$] or as a function of message format [$t(71) = 0.10$, $SD = 0.25$, $p = .92$]. Hence, the remaining analyses of the confidence-accuracy relationship were conducted on all of the recognition data pooled across the involvement and format conditions.

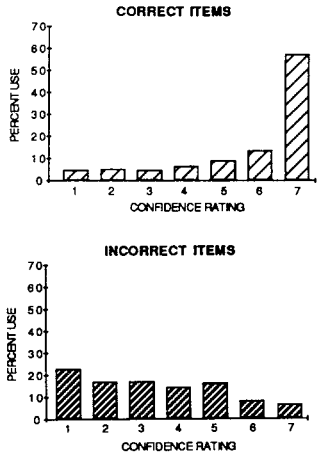

Figure 1. Contrast in confidence category usage as a function of recognition accuracy.

Accuracy and confidence category use. Figure 1 presents the distributions of confidence category use for correct and incorrect responses pooled across all of the subjects. Although the gamma correlations suggest that subjects as a group are quite sensitive to the likelihood that they have answered questions about the news broadcast correctly, an examination of the distribution of confidence ratings suggests that subjects are primarily sensitive to what they *do* know but not to what they *do not* know.

Notice that when the subjects were accurate, they were much more likely to report being *completely sure* (Rating 7) than they were to report having any other level of confidence. When the subjects were in error, however, they generally failed to differentiate among the lower confidence categories. Instead, their confidence ratings range indiscriminately between low and moderate values.

The distribution of confidence ratings when subjects are in error suggests that as a group subjects cannot accurately assess when they lack knowledge. Given this, the overall confidence-accuracy relationship may deteriorate as item difficulty increases because the ratio of errors to correct answers will increase. To examine this hypothesis, a post hoc analysis of item difficulty was performed.

Item difficulty and metamemory. For this analysis, the 40 items from the original memory test were reorganized according to subjects' likelihood of answering the items correctly. A preliminary correlation indicated a surprisingly strong relationship between mean item confidence ratings and item difficulty ($r = +.82$). The more difficult the item was to answer correctly, the lower was the mean confidence rating, despite the fact that the subjects had no way of knowing *a priori* which items would turn out to be most difficult to answer overall.

Although this relationship suggests that mean confidence ratings can be used to accurately predict which items are more likely to be answered correctly, gamma correlations were also computed to determine whether the confidence-accuracy relationship changes as a function of item difficulty. To do this, the original 40 items were separated into two groups based on the proportion of subjects who responded correctly to the items. The *easy* item group was made up of the 19 items with the highest accuracy rates (mean proportion correct = 83.2%), and the *hard* group consisted of the remaining 21 items (mean proportion correct = 57.6%). The gamma for the easy items ($\gamma = +.82$) was indeed significantly greater than the gamma for the hard items ($\gamma = +.62$) [$t(63) = 4.92$, $SD = 0.33$, $p < .001$], corroborating that there is a decline in the quality of the confidence-accuracy relationship as item difficulty increases.

The influence of item difficulty on the confidence-accuracy relationship is illustrated in Figure 2, which shows the differences in confidence category use across hard and easy items both for accurate performance and for errors. As can be seen in the graph at the top, the principal reason for the differences in confidence for accurate items is that subjects are less likely to report being completely sure of hard items. Notice, however, that their use of each of the remaining confidence categories increases at about the same rate, suggesting that when they are not completely sure they have trouble discerning which level of confidence is appropriate.

For errors, regardless of whether the items were hard or easy, subjects did not differ systematically in their use of the confidence categories. This is surprising in that one would expect that as items become more difficult, subjects should be more likely to realize that they have no idea what the correct answer is and to indicate this by increasing their use of the Rating 1 category (*not at all sure*). There is virtually no evidence here that this type of adjustment in category use occurs. Instead it seems that subjects cannot appropriately differentiate just how unsure they are of answers, only that they are not completely sure.

Predicting performance from confidence. Although the item difficulty analysis suggests that calibration will generally be worse for more difficult items, a correlation

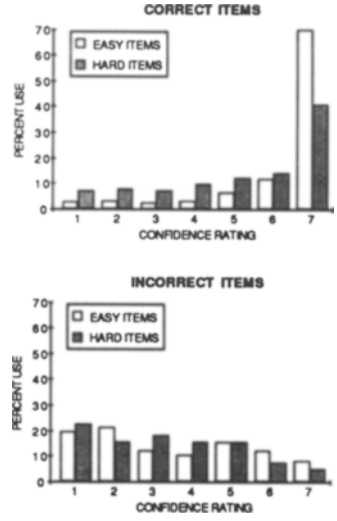

Figure 2. Confidence category usage as a function of item difficulty.

between each subject's performance accuracy and the corresponding gamma suggests that this relationship does not generalize to comparisons among individuals ($r = +.05$). Hence, subjects who perform exceptionally well on the test are not likely to be any better in the use of confidence ratings than are subjects who perform relatively poorly. Glenberg and Epstein (1987) contend that calibration scores may not correlate reliably with any other factor because individual calibration itself is not reliable. In their inference verification task, they found that individuals' gammas were not stable across testing sessions. If this instability generalizes to the present context, reliable gamma-accuracy relationships are not to be expected.[2]

Nevertheless, high scorers tend to maintain relatively high confidence levels and low scorers tend to maintain relatively low confidence levels ($r = +.67$). This demonstrates that one can reasonably predict which subjects have done well or poorly given only mean confidence ratings. What is especially interesting, though, is that high scorers tend to be more confident not only when they are correct, but also when they are incorrect. The correlation between subjects' average confidence ratings when their responses are correct and their average confidence ratings when in error is +.55. This may indicate that subjects experience a global sense of how much they remember from a particular broadcast, which serves as an anchor for making the more specific item judgments.

Awareness of idiosyncratic knowledge? A final test of subjects' awareness of their memory focuses on whether individuals have, as Nelson et al. (1986) put it, "privileged access to idiosyncratic knowledge." This test compares predictions of subjects' performance based on their own confidence ratings with predictions of their performance based on item difficulty (measured for the group as a whole). If individuals have special access to their knowledge, they should be able to predict their own performance more accurately than a predictor derived from group performance.

To examine this possibility, each subject's own confidence-accuracy gamma correlation was compared to a gamma correlation relating general item difficulty to the subject's recognition accuracy. Confidence values for the item difficulty (IDIF) gammas were constructed by assigning the easiest items (at the level of the group) with confidence ratings of 7 (*completely sure*) and the hardest items with confidence ratings of 1 (*not at all sure*), with intermediate items given intermediate ratings. The number of items assigned to each confidence category was determined by the number of items the subject had assigned to each category.[3] The computed IDIF gammas represent the relationship between these new IDIF confidence ratings and each subject's performance.

In contrast to Nelson et al.'s (1986) findings, the subjects' ability to predict their own performance (mean $\gamma = +.75$) was clearly superior to predictions based on item difficulty (mean $G = +.51$) [$t(71) = 8.44$, $SD = 0.25$, $p < .001$]. Sixty-two of the 72 subjects had higher gamma correlations based on their own confidence than based on item difficulty. At least in this task, it appears that subjects do have privileged access to idiosyncratic knowledge that allows them to predict their performance more accurately than others could using information regarding item difficulty.

DISCUSSION

In contrast to most assessments of the relationship between confidence and accuracy, this study found that the subjects' confidence ratings were highly predictive of their actual performance. The subjects in all conditions were remarkably well calibrated, regardless of the level of involvement or message format. Even though the level of calibration deteriorated slightly for difficult items, the subjects' individual confidence ratings were better predictors of their own performance than were predictions based on item difficulty. With regard to knowledge acquisition, high-involvement messages resulted in more accurate memory, but the editorial versus informational format manipulation had no impact on recognition performance.

Differences in Metamemory Across Tasks

Although the results suggest that metamemory in the context of listening to the news is quite good, there are several factors that preclude direct comparison with other findings. First, we selected a task that we had reason to suspect would result in some relationship between confidence and accuracy. As Glenberg and Epstein (1985) pointed out, this approach is in some ways closer to postdiction than prediction. That is, subjects only report their confidence after they have selected their answer. Hence, the type of metamemory we studied is highly constrained and will not necessarily generalize to other more difficult metamemory tasks.

Still, the results are impressive. Even when compared with other studies using comparable measures of calibration, our subjects' confidence-accuracy relationships were generally much stronger.[4] In their examination of the calibration of performance in a comprehension task, for instance, Glenberg and Epstein (1985, Experiment 3) found an average confidence-accuracy relationship of $r_{pb} = +.23$, whereas the average in this study was $\gamma = +.75$. In a related study, they reported an average calibration of performance of $\gamma = +.39$ (Glenberg & Epstein, 1987). Using more reliable multi-item tests, Pressley and Ghatala (1988) replicated the relatively weak confidence-accuracy relationships in reading comprehension, with subjects averaging only $\gamma = +.38$. Nevertheless, they did find performance awareness comparable to the present study in an analogy completion task (mean $\gamma = +.70$) and in a test of selecting word opposites (mean $\gamma = +.74$). Overall, it appears that our awareness of knowledge gained from listening to the news is at least as good as our awareness of other types of knowledge, and frequently may be better.

A second critical difference across tasks is item difficulty. As suggested in previous judgment research (Lichtenstein & Fischoff, 1977), calibration generally declines as item difficulty increases. Our own data sup-

port this conclusion (as do those of Pressley and Ghatala, 1988). Given this, one might posit that the good calibration found in our study is primarily the result of having test items that were not especially difficult. Indeed, the high levels of calibration found for the easy item set, with an average of 83% of subjects answering correctly, seem consistent with Lichtenstein and Fischoff's conclusion that calibration is generally best at roughly 80% accuracy levels.

Although the ease of items in this study may serve to inflate our assessment of the strength of the confidence-accuracy relationship overall, there still seems to be evidence that the relationship may be better for recognition of news items than for comparable assessments of reading comprehension. Glenberg and Epstein (1985, 1987), for example, found accuracy averages of between 72% and 79% in their calibration-of-performance task, which is actually slightly less difficult than the present task (with a mean of 70% accuracy). Given Lichtenstein and Fischoff's (1977) judgment findings, one might expect that Glenberg and Epstein's relatively easy task should yield high levels of calibration, but it does not. Also, reconsider Pressley and Ghatala's (1988) reading comprehension task, wherein accuracy averaged 61%, with a mean $\gamma = +.38$. The difficulty level is comparable to the present study's hard items, wherein accuracy averaged 58%. Nevertheless, the confidence-accuracy relationship for the latter more difficult items was substantially stronger with a mean $\gamma = +.62$. So, even with item difficulty taken into account, it appears that assessments of knowledge gained from listening to the news may be more accurate than assessments of knowledge gained from classroom reading comprehension exercises.

Admittedly, it may be that most metamemory paradigms are more demanding than the one used here. For instance, comprehension tasks require more than remembering which answer is correct; they also require that an inference be made. Calibration-of-judgment tasks (and some feeling-of-knowing tasks) involve accessing general world knowledge from semantic memory rather than merely recognizing information that has been presented quite recently.

The strength of the confidence-accuracy relationship may also be moderated by task constraints not present in this study. As discussed previously, calibration-of-comprehension tasks frequently require subjects to provide a confidence rating prior to viewing the question to be answered—a requirement that obviously makes accurate prediction more difficult. In feeling-of-knowing tasks, metamemory is measured only for the subset of items that is not recalled. This eliminates from consideration items that subjects do recall, and may consequently lead to underestimates of subjects' overall sensitivity to what they know. Although all confidence judgments seem similar in that subjects must judge whether they have access to knowledge that can reasonably assure a correct response to a particular question, the type of task and the way in which metamemory is measured undoubtedly affect the observed strength of confidence-accuracy relationships.

Understanding the Confidence–Accuracy Relationship

At least two aspects of the results of this study have implications for understanding the confidence-accuracy relationship within a broader framework. First, this study suggests that calibration will be reduced for difficult items because subjects have trouble differentiating levels of partial confidence. Second, the results suggest that subjects' global assessment of their task-related knowledge seems to permeate their confidence judgments for each individual item.

One of this study's most interesting findings is that subjects predominantly use the *completely sure* category to indicate their confidence when they are correct, but they do not regularly use the *not at all sure* category when they are incorrect. Instead, they tend to use the low and intermediate categories relatively indiscriminately. Even when the items are relatively difficult, subjects do not seem to recognize when they have no knowledge of the correct answer. Although they decrease their use of the *completely sure* category for these difficult items when they are correct, they distribute their use of the remaining confidence categories relatively evenly across the remaining partial confidence categories. When they are incorrect, on the other hand, they show only the slightest tendency to increase their use of the *not at all sure* category.

This pattern in the distribution of confidence category use suggests that subjects generally experience a relatively strong feeling of knowing when they are correct, but that they do not experience a complementary feeling of *not* knowing when they are incorrect. Moreover, when subjects are not completely sure, it appears that they cannot discriminate just how unsure they are. Instead, they seem to experience some sense of partial knowledge on most occasions when they are wrong.

From one item to the next, subjects show some sensitivity to item difficulty, but they also demonstrate a tendency to maintain similar confidence levels throughout. For example, subjects who have relatively high confidence when correct in their answers will also have relatively high confidence when their answers are incorrect. Although this could reflect a longstanding tendency to rely on some confidence categories more than others, it seems more likely (especially given that average confidence level is a relatively good predictor of actual performance) that this consistency reflects the incorporation of a global assessment of knowledge of the broadcast into the prediction of each unique item. This possibility is consistent with Glenberg's (1987) suggestion that subjects base their confidence judgments on their overall sense of familiarity with the topic at hand.

These conjectures also seem largely consistent with Gigerenzer, Hoffrage, and Kleinbolting's (1991) recently introduced Brunswikian theory of confidence. They argue that confidence judgments, in the absence of a strong

feeling of knowing (i.e., confidence < *completely sure*), are the consequence of inductive inference operating only on indirect cues to the answer. They suggest that confidence judgments in this case reduce to judgments of cue validity, which may themselves be hard to estimate particularly when questions are difficult or misleading.

It may be that when subjects are not completely sure of an answer, they only have a sense that they know something that is related to the answer, without being able to accurately judge the degree to which that related knowledge can assist them in isolating the correct answer. Whereas confidence may serve as a relatively good predictor of both individual performance and question difficulty, it also seems clear that we must be wary of "feelings of sort of knowing," because they generally produce overestimates of performance accuracy.

Knowledge Acquisition and Listening to the News

Beyond the confidence-accuracy relationship, this study also has implications for learning as a function of listening to the news. Not surprisingly, knowledge acquisition was better for high-involvement messages than for low-involvement messages. People remember information that affects their lives better than they remember information that does not, presumably because they process personally relevant information more extensively. Involvement seems a prime candidate for increasing learning in other contexts as well, including the classroom. Many teachers seem intuitively aware of the importance of involvement, trying to incorporate examples into their lectures in the hopes of increasing students' perceptions of the material's personal relevance.

Unlike involvement, message format did not affect knowledge acquisition. This may suggest that our format manipulation was weak, but it may also suggest that there is relatively little difference between editorials and informational news stories in practice. Although more passively stated, informational news items routinely present the same kinds of facts and opinions that persuasive editorial messages do. Nevertheless, because news items are generally construed as objective information, people may be more inclined to encode the information as fact rather than discounting it as mere opinion. Indeed, one of the current controversies in the area of mass communication involves the impact of a message's slant, or "frame," on the audience's perceptions of the news (Barkin, 1989; Gamson, 1989; Graber, 1989).

Conclusion

The results of this study are refreshing in that they seem so sensible. People remember more of a radio news broadcast when the messages are personally relevant. Whereas listening to the news is not an explicit learning task, people have a fairly good awareness of what they have just learned from the broadcast—although they could be more sensitive to what they did not learn.

REFERENCES

ANDREWS, J. C., & SHIMP, T. A. (1990). Effects of involvement, argument strength, and source characteristics on central and peripheral processing of advertising. *Psychology & Marketing*, **7**, 195-214.

APSLER, R., & SEARS, D. O. (1968). Warning, personal involvement, and attitude change. *Journal of Personality & Social Psychology*, **9**, 162-166.

ARKES, H. R., CHRISTENSEN, C., LAI, C., & BLUMER, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior & Human Decision Processes*, **39**, 133-144.

BARCLAY, C. R., & WELLMAN, H. M. (1986). Accuracies and inaccuracies in autobiographical memories. *Journal of Memory & Language*, **25**, 93-103.

BARKIN, S. M. (1989). Coping with the duality of television news: Comments on Graber. *American Behavioral Scientist*, **33**, 153-156.

BOWER, G. H., & CLARK, M. C. (1969). Narrative stories as mediators for serial learning. *Psychonomic Science*, **14**, 181-182.

BOWER, G. H., & WINZENZ, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, **20**, 119-120.

CRAIK, F. I. M., & LOCKHART, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, **11**, 671-684.

CRAIK, F. I. M., & TULVING, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, **104**, 268-294.

EINHORN, H. J. (1980). Overconfidence in judgment. In R. A. Schweder & D. W. Fiske (Eds.), *New directions for methodology of social and behavioral science: Fallible judgment in behavioral research* (pp. 1-16). San Francisco: Jossey-Bass.

EPSTEIN, W., GLENBERG, A. M., & BRADLEY, M. M. (1984). Coactivation and comprehension: Contribution of text variables to the illusion of knowing. *Memory & Cognition*, **12**, 355-360.

FISCHHOFF, B., SLOVIC, P., & LICHTENSTEIN, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 552-564.

GAMSON, W. A. (1989). News as framing: Comments on Graber. American Behavioral Scientist, **33**, 157-161.

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

GLENBERG, A. M., & EPSTEIN, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 702-718.

GLENBERG, A. M., & EPSTEIN, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, **15**, 84-93.

GLENBERG, A. M., SANOCKI, T., EPSTEIN, W., & MORRIS, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, **116**, 119-136.

GLENBERG, A. M., WILKINSON, A. C., & EPSTEIN, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, **10**, 597-602.

GOODMAN, L. A., & KRUSKAL, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732-764.

GRABER, D. A. (1989). Content & meaning. *American Behavioral Scientist*, **33**, 144-152.

HART, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning & Verbal Behavior*, **6**, 685-691.

HYDE, T. S., & JENKINS, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning & Verbal Behavior*, **12**, 471-480.

KATZ, E., ADONI, H., & PARNESS, P. (1977). Remembering the news: What the picture adds to recall. *Journalism Quarterly*, **54**, 231-239.

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

KORIAT, A., LICHTENSTEIN, S., & FISCHHOFF, B. (1980). Reasons for

confidence. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 107-118.

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

LICHTENSTEIN, S., & FISCHHOFF, B. (1980). Training for calibration. *Organizational Behavior & Human Performance*, **26**, 149-171.

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

LOVELACE, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 756-766.

MAHESWARAN, D., & MEYERS-LEVY, J. (1990). The influence of message framing and issue involvement. *Journal of Marketing Research*, **27**, 361-367.

MAKI, R. H., & BERRY, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 663-679.

MAKI, R. H., FOLEY, J. M., KAJER, W. K., THOMPSON, R. C., & WILLERT, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 609-616.

MORRIS, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 223-232.

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

MURPHY, A. H., & WINKLER, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, **79**, 489-500.

NELSON, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**, 109-133.

NELSON, T. O., & DUNLOSKY, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, **2**, 267-270.

NELSON, T. O., LEONESIO, R. J., LANDWEHR, R. S., & NARENS, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 279-288.

NELSON, T. O., LEONESIO, R. J., SHIMAMURA, A. P., LANDWEHR, R. F., & NARENS, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 279-288.

PETTY, R. E., & CACIOPPO, J. T. (1979a). Effects of forewarning of persuasive intent and involvement on cognitive responses and persuasion. *Personality & Social Psychology Bulletin*, **5**, 173-176.

PETTY, R. E., & CACIOPPO, J. T. (1979b). Issue-involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality & Social Psychology*, **37**, 1915-1926.

PETTY, R. E., & CACIOPPO, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality & Social Psychology*, **46**, 69-81.

PETTY, R. E., & CACIOPPO, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

PETTY, R. E., CACIOPPO, J. T., & GOLDMAN, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality & Social Psychology*, **41**, 847-855.

PETTY, R. E., CACIOPPO, J. T., & SCHUMANN, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, **10**, 135-146.

PRESSLEY, M., & GHATALA, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly*, **23**, 454-464.

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

SCHACHTER, D. L. (1983). Feeling of knowing in episodic memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 39-54.

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

STAUFFER, J., FROST, R., & RYBOLT, W. (1983). The attention factor in recalling network television news. *Journal of Communication*, **33**, 29-37.

TETLOCK, P. E., & KIM, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality & Social Psychology*, **52**, 700-709.

TOMASSINI, L. A., SOLOMON, I., ROMNEY, M. B., & KROGSTAD, J. L. (1982). Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior & Human Performance*, **30**, 391-406.

VESONDER, G. T., & VOSS, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory & Language*, **24**, 363-376.

WAGENAAR, W. A. (1969). Note on the construction of digram-balanced Latin squares. *Psychological Bulletin*, **72**, 384-386.

WEAVER, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 214-222.

NOTES

1. Mass communication researchers have generally concluded that broadcast news reports essentially go in one ear and out the other. In a study of recall of television news, Katz, Adoni, and Parness (1977) found that 21% to 34% of their subjects were unable to recall even one news item. The remaining subjects remembered an average of only 2 out of 14 or 15 items. In another representative study, Stauffer, Frost, and Rybolt (1983) telephoned some of their subjects the night before and asked them to watch their favorite national news program. Although this cued group recalled 58% more news items than did the noncued sample, neither group as a whole recalled more than 25% of the stories.

2. Due to several design differences, it is not entirely clear that Glenberg and Epstein's (1987) findings would generalize to the present context. In their study, gammas were based on a total of *16 true-false inference* items paired with confidence ratings made *prior* to viewing the question. In contrast, in the present study, gammas were based on a total of *40 recognition memory* items paired with confidence ratings made just *after* answering the question. This study's increase in item sample size and use of less cognitively complex tasks may improve the reliability of individual's gammas, relative to those found in Glenberg and Epstein (1987).

3. In a second set of IDIF gammas, the number of items assigned to each confidence category was determined by calculating the range of performance across all items (24 to 68 subjects correct = span of 45 "points") dividing it into seven roughly equal categories (either 6 or 7 points per category), and then assigning confidence categories to correspond optimally to these different levels of item difficulty. The procedure produced only minor differences in the obtained gamma correlations, and the results were entirely consistent with what is reported for the original IDIF gammas based on each subject's own distribution of confidence ratings.

4. This is all the more impressive when one considers that a pilot study was conducted to remove very easy and very difficult recognition items from the final question set. The net result is a more restricted range of test items, which should presumably be harder to differentiate than a set of items that spanned the entire continuum.

(Manuscript received April 15, 1992;
revision accepted for publication August 21, 1992.)