METAMEMORY: A THEORETICAL FRAMEWORK AND NEW FINDINGS

Thomas O. Nelson
Louis Narens

I. Introduction

Although there has been excellent research by many investigators on the topic of metamemory, here we will focus on our own research program. This article will begin with a description of a theoretical framework that has evolved out of metamemory research, followed by a few remarks about our methodology, and will end with a review of our previously unpublished findings. (Our published findings will not be systematically reviewed here; instead, they will be mentioned only when necessary for continuity.)

II. A Theoretical Framework for Metamemory

A. Three Abstract Principles of Metacognition

Our analysis of metacognition is based on three abstract principles that have been individually used in isolation by other authors:

Principle 1: The cognitive processes are split into two or more specifically interrelated levels. Figure 1 shows the basic structure, which contains two interrelated levels that we call the *meta-level* and the *object-level,* following the usage of those terms by the mathematician Hilbert

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 26

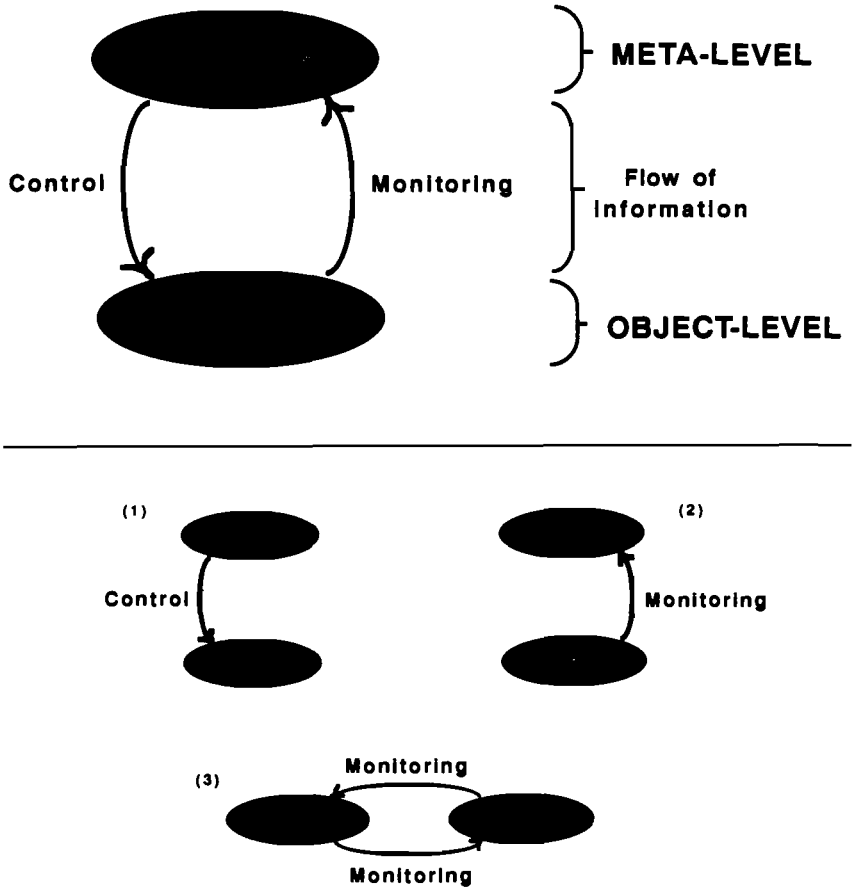

Fig. 1. Upper panel shows a theoretical mechanism consisting of two structures (meta-level and object-level) and two relations in terms of the direction of the flow of information between the two levels (notice the asymmetric aspect of each relation). Lower panel shows (1) a nonhomeostatic mechanism without any feedback, (2) a spylike mechanism that has information about the system but no control (e.g., a time traveller who isn't allowed to affect history), and (3) a mechanism with a symmetric relation, such that neither component is meta-level with regard to the other (e.g., two department chairmen discussing their respective departments).

(1927; i.e., "metamathematics") and by the philosopher Carnap (1934; i.e., "metalanguage"). Generalizations to more than two levels can be developed, but we have no need to do so for this article.

Principle 2: The meta-level contains a dynamic model (e.g., a mental simulation) of the object-level. Conant and Ashby (1970) gave a demon-

stration for the necessity of such an assumption if the system is to control a dynamic process so as to change from a given state to some other goal state.

Principle 3: There are two dominance relations, called "control" and "monitoring," which are defined in terms of the direction of the flow of information between the meta-level and the object-level. This distinction in the direction of flow of information is analogous to that in a telephone handset, as discussed next.

1. Control

The basic notion underlying control—analogous to speaking into a telephone handset—is that the meta-level *modifies* the object-level. In particular, the information flowing from the meta-level to the object-level either changes the state of the object-level process or changes the object-level process itself. This produces some kind of action at the object-level, which could be (1) to initiate an action, (2) to continue an action (not necessarily the same as what had been occurring because time has passed and the total progress has changed, e.g., a game player missing an easy shot as the pressure increases after a long series of successful shots), or (3) to terminate an action. However, because control per se does not yield any information from the object-level, a monitoring component is needed that is logically (even if not always psychologically) independent of the control component.

2. Monitoring

The basic notion underlying monitoring—analogous to listening to the telephone handset—is that the meta-level is *informed* by the object-level. This changes the state of the meta-level's model of the situation, including "no change in state" (except perhaps for a notation of the time of entry, because the rate of progress may be expected to change as time passes, e.g., positively accelerated or negatively accelerated returns). However, the opposite does not occur, i.e., the object-level has no model of the meta-level. The main methodological tool for generating data about metacognitive monitoring consists of the person's subjective reports about his or her introspections.

3. Role of Subjective Reports about Introspection for Inferences about Monitoring

During the past decade or so, subjective reports about introspection have been resurrected in a form that circumvents the serious flaws in the

older version used by turn-of-the-century psychologists. Methodological rigor is increased when people are construed as imperfect measuring devices of their own internal processes and when the assumption that introspection yields a veridical picture of the person's internal processes is not made. This distinction in our use of subjective reports is critical and can be highlighted by noticing an analogy between the use of introspection and the use of a telescope. One use of a telescope (e.g., by early astronomers and analogous to the early use of introspection) is to assume that it yields a perfectly valid view of whatever is being observed. However, another use (e.g., by someone in the field of optics who studies telescopes) is to examine a telescope in an attempt to characterize both its valid output and its distortions. Analogously, introspection can be examined as a type of behavior so as to characterize both its correlations with some objective behavior (e.g., likelihood of being correct on a subsequent test) and its distortions.

Thus we try to recognize and avoid the potential shortcomings of introspection (e.g., Nisbett & Wilson, 1977) while capitalizing on its strengths (e.g., Ericsson & Simon, 1980, 1984). We view introspective reports as data to be explained, in contrast to the Structuralists' view of introspective reports as descriptions of internal processes; i.e., we regard introspection not as a conduit to the mind but rather as a source of data to be accounted for by postulated internal processes.

Although previous writers such as Nisbett and Wilson (1977) have underscored the possibility of distortions in introspective monitoring, they have not emphasized its potential role—even with its distortions—of affecting control processes. *A system that monitors itself (even imperfectly) may use its own introspections as input to alter the system's behavior.* One of our primary assumptions is that in spite of its imperfect validity and in spite of its being regarded by some researchers as only an isolated topic of curiosity, introspection is a critical component in the total memory system. In attempting to understand that system, we examine the person's introspections so as to have some idea about the input that the person is using.

The person's reported monitoring may, on the one hand, miss some aspects of the input and may, on the other hand, add other aspects that are not actually present. Indeed, one of our goals is to characterize both the accuracy and the distortions that are present in people's introspections. This is analogous to one traditional view of perception, where what is perceived is different from what is sensed (i.e., perception conceptualized as sensation plus inference), except that what is analogous to the objects being sensed is here the object-level memory processes.

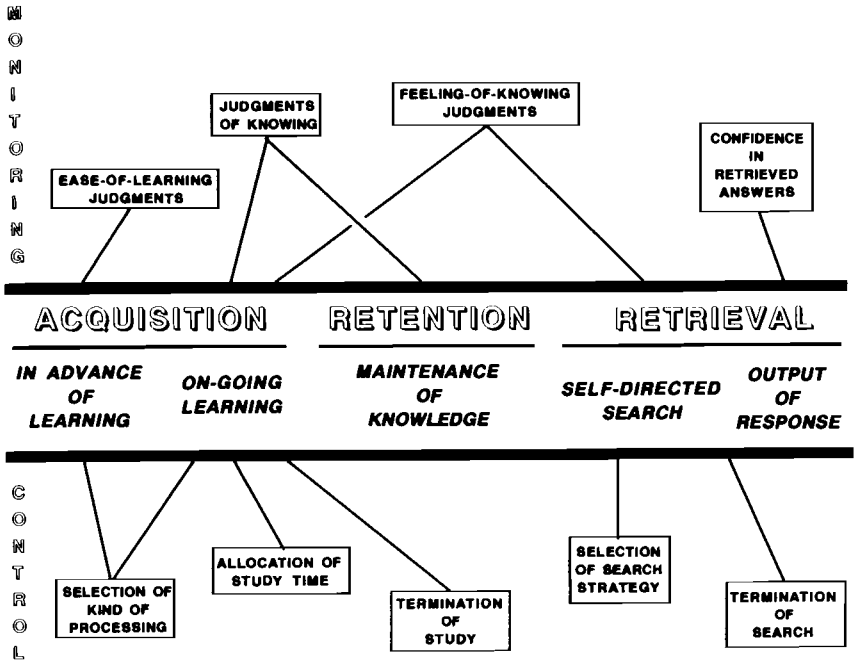

Fig. 2. Main stages in the theoretical memory framework (listed inside the horizontal bars) and some examples of monitoring components (shown above the horizontal bars) and control components (shown below the horizontal bars).

B. The Monitoring and Control of Human Memory

An overview of our theoretical framework is shown in Fig. 2. The monitoring and control processes are grouped in terms of the overall stages of the system, as discussed next, and the reader is invited to consider them in the context of a college student studying for an upcoming examination.

1. Acquisition Stage: In Advance of Learning

Two components that occur in advance of learning consist of the person's goal and the person's plan to achieve that goal.

a. Determining One's Goal: The Person's Norm of Study. When the person becomes aware of the to-be-remembered items and the anticipated type of test, he or she makes a judgment about the level of mastery that will be needed for a given item at the time of the anticipated test. When

a delay is expected to occur between acquisition and the retention test, then the person's *theory of retention* (Maki & Berry, 1984) is used to modulate how well each item would have to be mastered now, in order for it to still be remembered on the retention test. The product—of the desired ease of retrieval during the retention test, modulated upward by however much extra learning the person believes will be needed to breach the retention interval—is the overall degree of mastery the person believes should be attained during acquisition, which is referred to as the person's *norm of study* (Le Ny, Denhiere, & Le Taillanter, 1972).

b. Formulating a Plan to Attain the Norm of Study. After the norm of study has been determined, the person makes a decision about how to attain that goal, i.e., formulates a plan. This has several parts, involving several kinds of monitoring judgments that need to be distinguished.

First, a distinction should be drawn between retrospective monitoring (e.g., a confidence judgment about a *previous* recall response) vs. prospective monitoring (e.g., a judgment about *subsequent* responding). The latter are subdivided further into three categories in terms of the state of the to-be-monitored items:

1. Ease-of-learning (EOL) judgments occur in advance of acquisition, are largely inferential, and pertain to items that have not yet been learned. These judgments are predictions about what will be easy/difficult to learn, either in terms of which items will be easiest or in terms of which strategies will make learning easiest.
2. Judgments of learning (JOL) occur during or after acquisition and are predictions about future test performance on currently recallable items.
3. Feeling-of-knowing (FOK) judgments occur during or after acquisition (e.g., during a retention session) and are judgments about whether a given currently nonrecallable item is known and/or will be remembered on a subsequent retention test.

Perhaps surprisingly, EOL, JOL, and FOK are not themselves highly correlated (Leonesio & Nelson, 1990). Therefore, these three kinds of judgments may be monitoring somewhat different aspects of memory, and whatever structure underlies these monitoring judgments is likely to be multidimensional (speculations about some possible dimensions occur in R. Krinsky & Nelson, 1985, and Nelson, Gerler, & Narens, 1984, esp. pp. 295–299).

c. Ease-of-Learning Judgments. Initially, the person makes and EOL judgment about the degree of difficulty for each item (or set of items) in terms of acquiring that item to the degree of mastery set by the

norm of study. Underwood (1966) showed that EOL is an accurate predictor of the rate of learning during experimenter-paced study trials, and we showed that EOL is related to how much study time is allocated to each item during self-paced study trials (Nelson & Leonesio, 1988, discussed below).

d. A Priori Choice-of-Processing Judgments. After making EOL judgments, the person decides which of the various kinds of processing to use on the to-be-retrieved items, and this decision can affect the rate of learning.

e. Initial Plan for the Allocation of Study Time. When planning the allocation of study time, the person may first determine the total time to allocate (e.g., 4 hr of study for an upcoming exam). The kind of retention test that is anticipated may affect both the planned allocation of self-paced study time and how that self-paced study time is apportioned among the items (Butterfield, Belmont, & Peltzman, 1971), including massed vs. distributed self-controlled rehearsals (Modigliani & Hedges, 1987).

We investigated the relation between EOL and the allocation of self-paced study time, with the major finding being that people study longer on the items they believe in advance will be harder (Nelson & Leonesio, 1988). The specific model explored in that research is reproduced here in Fig. 3, both to illustrate how hypothetical causal relations between monitoring and control processes can be explored and to show how the theoretical constructs of the framework can be operationalized, with monitoring constructs typically being operationalized via an introspective report (e.g., EOL judgment) and control constructs being operationalized by some other empirical outcome (e.g., elapsed time during self-paced study).

2. *Acquisition Stage: The Ongoing Learner*

The focus here is on the changes in both the learner's plan and the learner's performance. Figure 4 shows a model of some hypothetical causal relations between several metacognitive components during the ongoing aspects of acquisition. This model may be useful both as a guide to the components discussed here and as an example of one way that stronger models can be developed within our framework. The metacognitive components contained in the model are shown in the upper portion of Fig. 4; the lower portion includes a basic memory model (cf. Atkinson & Shiffrin, 1968; Ericsson & Simon, 1980), containing a working memory (cf. short-term memory, STM) that is separate from long-term memory (LTM).

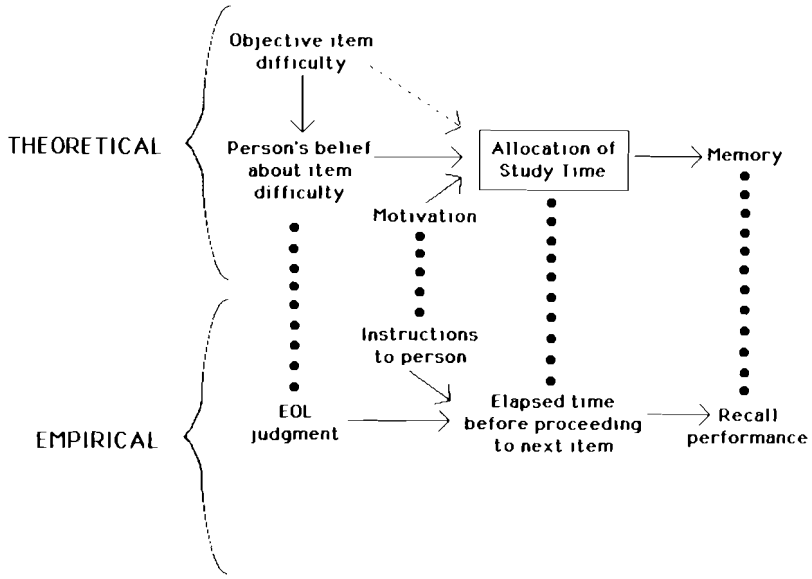

Fig. 3. A model of the allocation of self-paced study time, with arrows indicating hypothesized causal connections and dotted lines indicating the way in which each theoretical construct is operationalized (after Nelson & Leonesio, 1988).

The upper left corner of Fig. 4 shows three metacognitive components discussed earlier that give rise to the person's norm of study. Following attempts to learn a given item, a judgment is made about the current state of mastery for that item (namely, an FOK judgment if the item is not currently recallable, or a JOL if the item is currently recallable). When the current state of mastery reaches the norm of study, the person terminates study of that item (i.e., exits from the sequence). However, when the current state of mastery has not reached the norm of study, the person allocates more study time to the item, chooses a strategy from his metacognitive library of strategies (which may reside in a portion of permanent memory—not shown in Fig. 4; cf. the concept of "metacognitive knowledge" in Flavell, 1979), and implements the strategy in an attempt to attain the desired degree of mastery for that item. Then the cycle recurs. Each of these hypothesized aspects of the ongoing cycle is elaborated below.

a. Feeling of Knowing for Currently Nonrecallable Items. Nelson and Leonesio (1988, Experiment 3) found that FOK judgments made after failed attempts at recall of general-information items were positively cor-

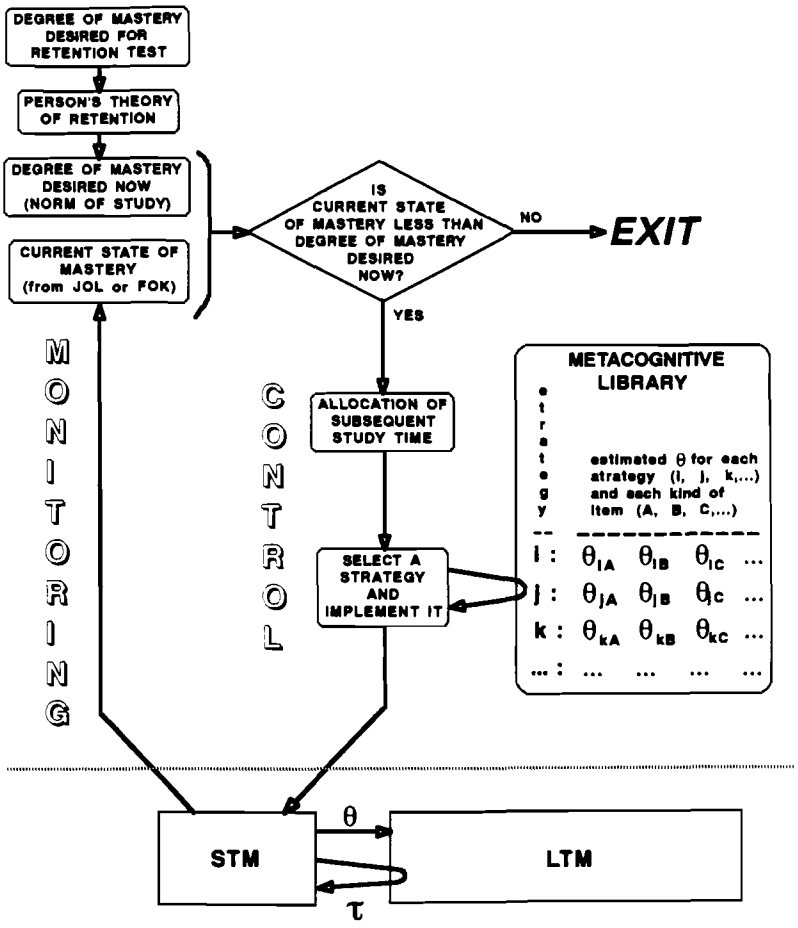

Fig. 4. Example of some metamemory components (shown above the dotted line) during acquisition. Curved-return arrow indicates that a given component obtains information from another component. Information is acquired into LTM at a rate θ and is retrieved from LTM with probability τ (Atkinson & Shiffrin, 1968; other views of storage and retrieval, including the substitution of working memory for STM, could easily be substituted instead). The metacognitive library tells for each available strategy (i, j, k, . . .) the estimated rate of acquisition it will yield for various kinds of items (A, B, C, . . .).

related with the subsequent allocation of self-paced study time on those items (cf. Figs. 3 and 4 above). However, as in the case of EOL, the magnitude of that correlation was far from unity, indicating that additional mechanisms underlie the person's allocation of study time (discussed below).

b. Judgments of Learning for Currently Recallable Items. According to Ericsson and Simon (1980), the monitoring per se occurs in STM. This does not, however, imply that information in LTM cannot be monitored—e.g., people are aware that they know their own names. Information that is in LTM may be monitored by first copying it into a working memory, also referred to as STM (cf. lower portion of Fig. 4), such that the person can functionally monitor both STM and LTM (latencies of that monitoring are reported in Wescourt & Atkinson, 1973). Unfortunately, however, people may mistakenly assess their JOL by monitoring information that is only in STM, not in LTM. When that occurs, the JOL predictions are likely to be accurate for predicting subsequent short-term recall of that information but may be inaccurate for predicting subsequent long-term recall (e.g., on a later examination). Would it be possible to produce more accurate JOL predictions for subsequent long-term recall if people made their JOL after a brief delay from when a given item was studied, so as to minimize recall of the to-be-judged information from STM and instead require recall from LTM? We have begun research on this topic, and preliminary results (J. Dunlosky & T. O. Nelson, unpublished) indicate that the answer to this question is affirmative.

c. Updating the Allocation of Study Time during a Particular Study Trial of an Item. In contrast to Fig. 3 but as indicated in Fig. 4, there may be an ongoing allocation of study time to an individual item in the list, such that the person continues studying until his or her JOL for the item reaches the norm of study. The circumstances under which people sharpen their differential study time (i.e., devote much more study time to harder items and much less study time to easier items) have not yet been established (but see Nelson & Leonesio, 1988; Mazzoni, Cornoldi, & Marchitelli, 1990).

Related to this, we found that college students terminate self-paced study on a given item long before it has been mastered well enough for subsequent recall (Nelson & Leonesio, 1988). For instance, in our Experiment 1, people who were specifically instructed to continue the self-paced study of each item until they were sure that they would be 100% correct on an upcoming recall test ended up having only 49% correct recall when tested after the study phase. That research examined only one self-paced study trial per item. Future research should determine whether

the same or different results occur during multitrial acquisition, because people routinely learn information to mastery, and this needs to be reconciled with the Nelson and Leonesio findings. The role of motivation in allocating study time should also be explored more fully.

d. Termination of Acquisition. How does a learner decide when to terminate acquisition? Figure 4 can be regarded as one answer to this question, with termination occurring when the JOL reaches the norm of study, as discussed above.

3. *Retention Stage*

The major metacognitive activity during this stage is the maintenance of previously acquired knowledge (see Bahrick & Hall, 1990). Several factors may underlie the person's decision about how and when to review. For instance, the person may have a theory of forgetting that includes the hypothesis (empirically confirmed by Leonesio & Nelson, 1982) that the hardest item to learn will be the hardest item to retain.

People potentially could capitalize on their metacognitive monitoring of items to decide how much subsequent study to devote to various items that cannot be recalled on a given maintenance test. Perhaps the mechanism would be similar to that for acquisition, where additional processing of a given to-be-retrieved item depends upon the discrepancy between the desired degree of mastery for the item vs. the assessed degree of mastery (cf. Fig. 4). For nonrecallable items, the person's FOK may help to direct whatever maintenance—more aptly, "relearning" for nonrecallable items—is allocated (Nelson *et al.*, 1984, Experiment 1; Nelson & Leonesio, 1988, Experiment 3).

4. *Retrieval Stage: Termination*

Nickerson (1980) distinguished between memory retrieval that is versus is not self-directed. Although knowledge about both kinds of retrieval is important for memory theory, our framework focuses on self-directed retrieval. The self-direction occurs not in the searching itself (which we assume to be automatic once it is initiated—see Fig. 5), but rather in setting up the particular cues to initiate the search (e.g., by consciously thinking of the last episode in which the item was retrieved or by consciously going through the alphabet as cues for the first letter of the sought-after answer).

Some components we suppose are involved in the termination of the retrieval stage are shown in Fig. 5, which shows mechanisms for continuing vs. terminating the stage of memory retrieval.

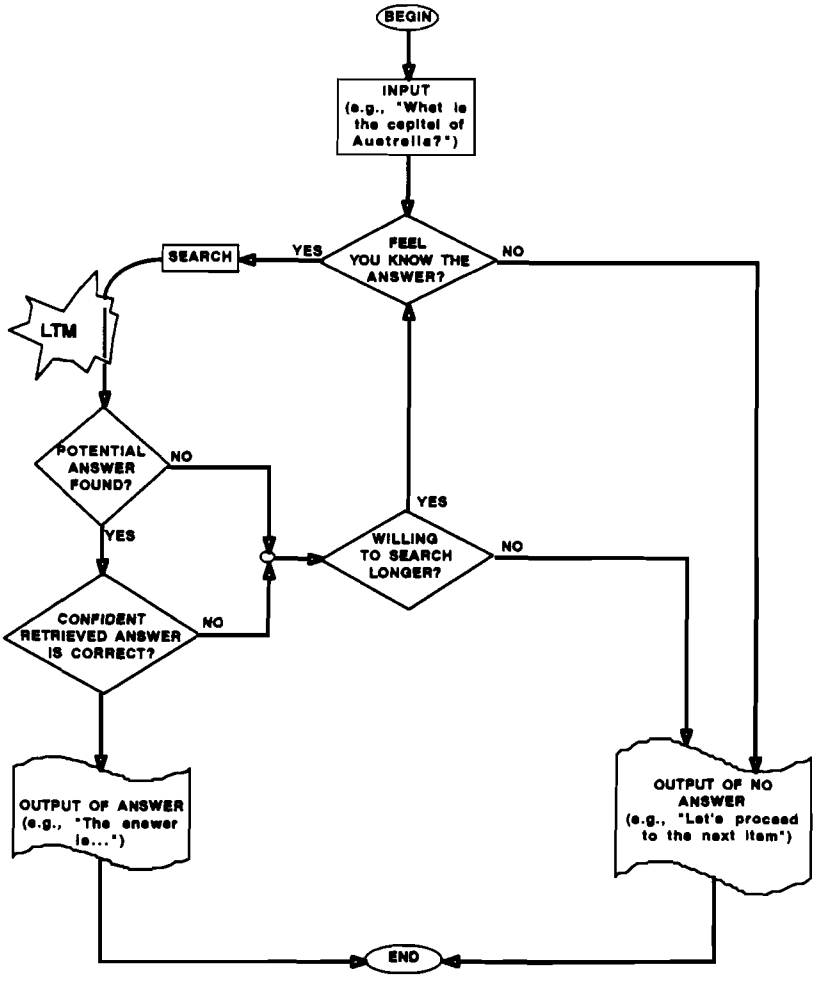

Fig. 5. Some metamemory components in the retrieval stage in human memory.

a. Quick Initiation/Termination of Retrieval. The metacognitive decision to initiate a search appears to be based on a very rapid, preliminary FOK judgment (Reder, 1987, 1988). This may be similar to the decision that people make in television game shows such as "Jeopardy" that require the player to signal rapidly that he or she can answer a given question. Upon presentation of a general-information question, people can make fairly accurate FOK judgments (about whether or not they could recall the answer) with a latency that is shorter than the latency of actu-

ally recalling the answer. Accordingly, in Fig. 5 this preliminary FOK judgment precedes recall (see Schreiber, Nelson, and Narens' research discussed in Section III, B, 7, a).

This mechanism may be similar to the one postulated in Juola *et al.*'s (1971) model for a "fast yes" or "fast no" response in yes/no recognition. The "fast no" may be based on the person's belief of never having encountered the requested information, as in Kolers and Palef's (1976) "knowing not" (also see Nelson *et al.*, 1984, p. 297, for the role of memory for prior encounters as a basis for FOK.)

b. Placement of Retrieval-Termination Threshold for Nonretrieved Items. As indicated in Fig. 5, when a potential answer is not found on a given search through memory, people presumably make a decision about whether they are willing to expend more time searching for the answer (i.e., using some kind of costs/rewards rules). If they are, and if the FOK is still positive, the search continues. However, the FOK may no longer be positive enough to continue. That is, there may be an evaluation of progress that is dynamic for a given item (e.g., an evaluation in terms of whether there has been sufficient progress to continue). When someone either is no longer willing to continue searching for the item or has a reduced FOK that no longer exceeds the FOK threshold for claiming to know the answer, the process is terminated with an omission error (indicated in the right-hand side of Fig. 5). The relationship between FOK and how long the retrieval stage continues prior to an omission error has been established empirically: Greater FOK is correlated with a longer latency of an omission error (Nelson *et al.*, 1984, Fig. 3).

The aforementioned mechanisms for terminating searching should be distinguished sharply from the ones in the left-hand side of Fig. 5, where a potential answer is retrieved and output. Then when the outputed answer is incorrect—i.e., a commission error—the relationship between FOK and the latency of that error is nil (Nelson *et al.*, 1984, Fig. 3). Commission-error latencies probably involve a complicated mix of confidence judgments and other factors (discussed below). Moreover, people's FOK is not completely accurate and is sometimes mistaken because they retrieve the wrong referent (e.g., retrieving Sydney in response to the question, "What is the capital of Australia?"; R. Krinsky & Nelson, 1985; also see Schacter & Worling, 1985).

Omission versus commission errors have also yielded different effects on other aspects of metacognition. For instance, college students typically have a greater FOK for commission-error items than for omission-error items, even though there is no difference in subsequent recognition memory on the two kinds of items (R. Krinsky & Nelson, 1985).

The person's expected reward for correct retrieval can affect the decision to continue or terminate searching (i.e., the threshold for "willingness to search longer" in Fig. 5). Although incentive can affect how long the person will continue before terminating the retrieval stage (Loftus & Wickens, 1970), there is no empirical evidence about whether greater incentive can produce a greater probability of retrieving during a given amount of retrieval time.

5. *Retrieval Stage: Output of Response*

Several potential psychological mechanisms may underlie the decision to output a single retrieved answer. Some versions of generation-recognition models of recall (e.g., Bahrick, 1970) propose that a "recognition stage" occurs in which the person makes a yes/no recognition judgment and on that basis decides whether to output the answer that he or she retrieved (i.e., "generated"). If the person retrieves only one response that seems plausible, then presumably that response is evaluated against a confidence threshold like the one indicated in Fig. 5 (confidence judgments per se are discussed in the next section). Perhaps this process is mediated by some kind of conscious recollection.

A variant of the aforementioned mechanism is what might be labeled the test-until-deemed-successful strategy: If the amount of confidence for the first answer that the person retrieves is below the confidence threshold, and if the person continues to search but does not retrieve any other potential answers for that item, then the confidence threshold might be lowered (i.e., a dynamic process). Accordingly, the initially retrieved answer might be output even though it was not associated with enough confidence to be output earlier.

Another strategy can occur in which people output an answer even when they are not convinced that it is correct, but rather only that it has a good likelihood of being correct. This satisficing strategy consists of "aiming at the good when the best is incalculable . . . some stop rule must be imposed to terminate problem-solving activity. The satisficing criterion provides that stop rule: retrieval ends when a good-enough alternative is found" (Simon, 1979, p. 3).

Any of the aforementioned strategies may also be modulated by external factors. For instance, the person's threshold for outputing a retrieved answer might be affected by the costs vs. rewards associated with commission vs. correct responses and/or might also be affected by drugs. The likelihood of commission errors during recall is known to increase after the ingestion of marijuana (Hooker & Jones, 1987; Pfefferbaum, Darley, Tinklenberg, Roth, & Kopell, 1977) or lithium (Weingartner, Rudorfer, & Linnoila, 1985) but is unaffected by alcohol (Nelson, McSpadden,

Fromme, & Marlatt, 1986). Although marijuana affects the threshold for outputing retrieved answers, it has no effect on the probability of correct recall (or on the FOK threshold for saying that correct recognition would occur). Nelson *et al.* (1990) found a related outcome at Mount Everest: High altitude decreased the likelihood of commission errors without affecting the probability of correct recall.

6. *Retrieval Stage: Confidence Judgments after Recall*

The confidence judgments that occur after the recall that the confidence pertains to are interesting, but their interpretation is difficult because they are validated retrospectively (in contrast to monitoring judgments such as EOL, JOL, and FOK, all of which are validated prospectively).

The usual finding is that people are overconfident—in terms of absolute scales—about their preceding memory performance (for a review, see Lichtenstein, Fischhoff, & Phillips, 1982), and this finding occurs across a wide variety of conditions (e.g., the degree of overconfidence about recall is approximately the same for normal and alcohol-intoxicated people; Nelson, McSpadden *et al.*, 1986). However, sometimes near-perfect calibration does occur (e.g., Nelson *et al.*, 1990).

Moreover, the reported confidence about the likelihood of an outputed answer being correct is not necessarily a direct measure of the person's internal confidence. For instance, when the person has retrieved two plausible answers for an item—each of which is associated with high internal confidence—but the experimenter allows only one to be output, the reported confidence may be low (because of the person's awareness that the other answer may be the correct one). However, if subsequently the person receives feedback that the outputed answer was wrong, then he or she may give a high FOK judgment—in contrast to the aforementioned low confidence judgment—because of the belief that the remaining answer must be correct. Accordingly, the probability of reminiscence (i.e., correct recall on a subsequent test after incorrect recall of the answer on a previous test, without any intervening study of the answer) is greater for commission-error items ($p = .29$) than for omission-error items ($p = .05$), perhaps because during the original test the person may sometimes think of two possible answers (one of which is correct and the other of which is incorrect), output the incorrect one, and then output the other one on a subsequent test of that item (Nelson *et al.*, 1984).

C. Refining the Components of the Theoretical Framework

Earlier we mentioned that Fig. 2 showed only a skeleton of our framework. Fleshing out that skeleton can be accomplished by what is referred

to in set-theory terminology (e.g., Shafer, 1976) and computer-assisted-design (CAD) terminology (e.g., Snow, 1987) as "coarsening" and "refining."

Within our framework, a coarsened node is a node that is elaborated at a greater level of specificity (i.e., containing more detail) somewhere else in the framework, and this elaboration at a greater level of specificity is the refinement. The key idea is that larger or smaller degrees of specificity can occur for any component of interest in the framework. To illustrate how refinement can occur, Fig. 6 shows the relation between the coarsened nodes of "Termination of Study" and "Termination of Search" that appeared in Fig. 2 and their refinement that appeared in Figs. 4 and 5, respectively.

Notice an important characteristic of this approach to theorizing: There is no need to preestablish any primitives at a "lowest level of specificity" or even to speculate about what the lowest level of specificity might be like. Also, a particular refinement may for convenience be represented as a coarsened node when it appears in the refinement of still other nodes.

III. Methodology and New Findings from Our Research

At the outset of our research on metamemory (circa 1975), we were aware of two findings about metamemory that seemed paradoxical in terms of the then-prevailing theories of memory. The first finding (Hart, 1965) was that individuals who failed to recall answers to general-information questions can nevertheless evaluate whether or not they know the answer, and such evaluations are positively correlated with their performance on a subsequent recognition test of the previously nonrecalled items. How could subjects consciously and validly monitor such nonrecallable answers?

The second finding (Juola, Fischler, Wood, & Atkinson, 1971; Kolers & Palef, 1976) was that subjects are able to give very fast and valid "No" answers to questions about whether they could recall specific items, and these "No" answers occur more quickly than a search of memory (i.e., a search of memory would take more time than the latency of the "No" response; for direct evidence that the latencies of FOK judgments are shorter than the latencies to search memory, both for "No" responses and for "Yes" responses, see Reder, 1987, 1988). How could a subject know about the presence/absence of an item in memory without first completing the memory search for it?

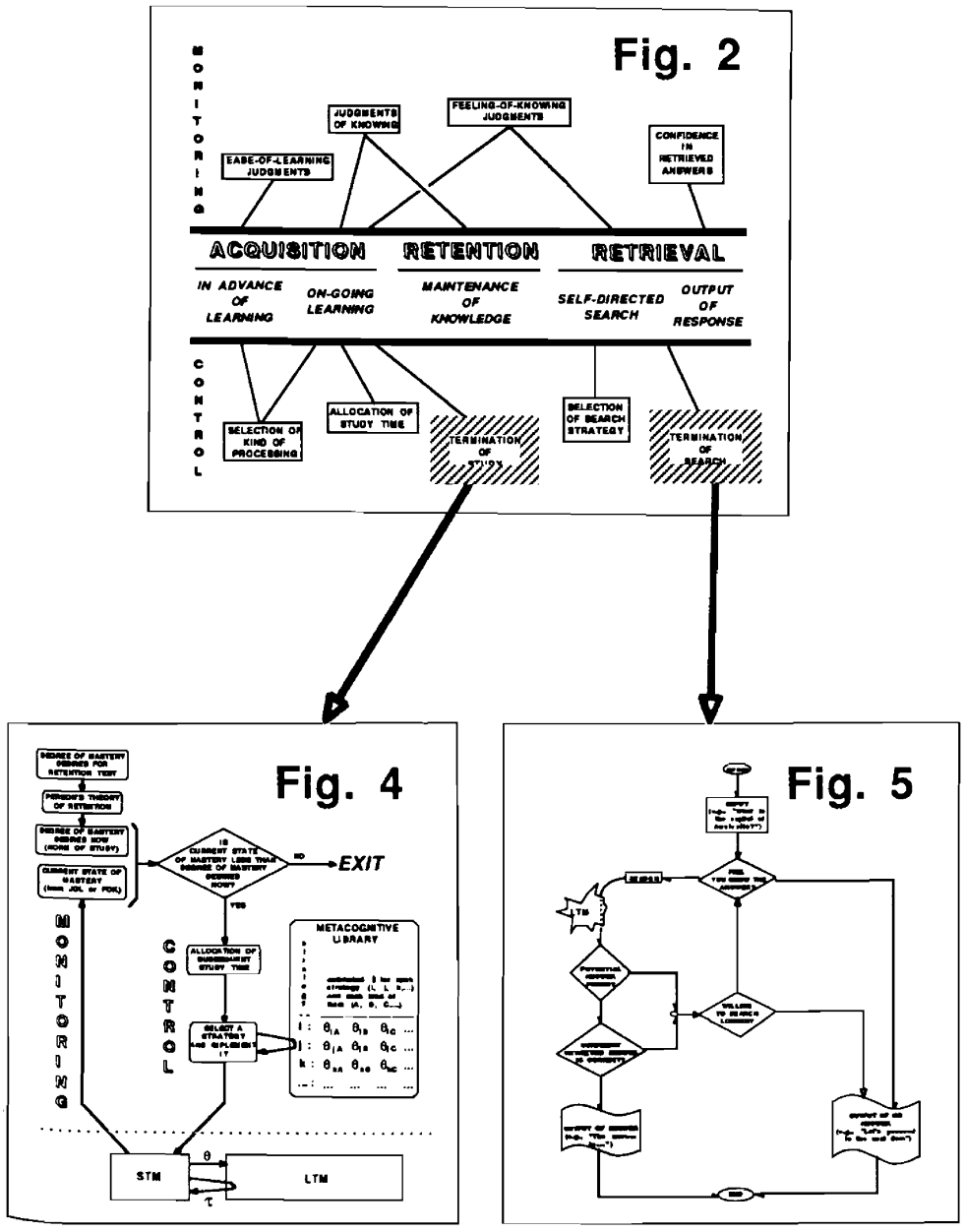

Fig. 6. Current status of refinement of coarsened nodes in the theoretical framework (see text for explanation of coarsening/refining).

A. Remarks about Our Methodology

We were intrigued by these paradoxes and decided to develop a research program to investigate metamemory experimentally. Analogous to the traditional psychophysical/measurement techniques in which people are construed as measuring devices of external stimuli, our approach to metamemory research was to construe people as measuring devices of their own *internal* stimuli. We hoped that this would allow us to determine how people monitored their own object-level cognitions (when we compared their judgments with our own assessments of their object-level cognitions, so as to see distortions) and also might give us information about their object-level cognitions that our assessments did not show. Accordingly, we based our methodology on obtaining, scaling, and comparing introspective judgments similar to those from psychophysical paradigms. However, the paradigm had to be modified both because of theoretical considerations and because of various practicalities (e.g., in metamemory experiments, the experimenter collects relatively few observations—with respect to the usual psychophysical case—from relatively many people—again with respect to the usual psychophysical case). We also had a major problem to deal with (as do the researchers in psychophysics and in recognition memory; see Shepard, 1967), namely, to avoid confounding two aspects of the person's judgments: (1) accuracy of the judgment (e.g., as had been assessed by d' in psychophysics) vs. (2) placement of the decision threshold for making the judgment (e.g., as had been assessed by β in signal-detection theory); for our investigations of metamemory, we sought both an a priori solution via new data-collection techniques and an a posteriori solution via new data-analysis techniques that led us to consider alternatives to d'.

1. Remarks about Data Analysis for Metacognitive Judgments

Although the signal-detection measure of sensitivity d' had been a useful statistic to compare performance across individuals and conditions in psychophysics, for various reasons (elaborated in Nelson, 1984, 1987; Nelson, McSpadden *et al.*, 1986) we chose to use Goodman and Kruskal's gamma, G, as the measure of metacognitive accuracy. In contrast to d', no distributional assumptions need to be made for G (such assumptions are critical for the use of d' and are untestable in most metamemory situations; for relevant discussions, see Lockhart & Murdock, 1970; Nelson, McSpadden *et al.*, 1986). In contrast to other correlations such as Pearson r or Spearman rho, G is unaffected by ties, which are unavoidable in metamemory research and are otherwise problematic. Also, the expected value of G is constant across changes in the person's

threshold for being confident. Finally, G has a very general interpretation in terms of telling the probability of accurate detection (see Nelson, 1984, Eq. 7; 1987, p. 305; Nelson, McSpadden *et al.*, 1986, Eq. 1), which yields a quantitative metric for the degree of FOK accuracy that is both intuitive and superior to any comparisons of difference scores (e.g., as in Hart, 1965, 1967; see Nelson, 1984, for reasons).

After evaluating all of the available measures, we concluded that G is the best measure of detection accuracy for research on metacognition (for more recommendations regarding the use of G, see Nelson, 1984, 1986).

2. *Remarks about Data Collection for Metacognitive Judgments*

a. Ratings vs. Rankings. One way to obtain people's metacognitive confidence judgments is to collect confidence ratings on an M-place Likert scale about the person's subjective impressions on each item (where $M \geq 2$). Then the validity of those judgments can be determined either (1) by plotting a calibration curve to determine the accuracy of absolute confidence (i.e., confidence for a given item relative to the person's threshold for being confident; Lichtenstein *et al.*, 1982) or (2) by computing G to determine the accuracy of relative confidence (i.e., confidence for one item relative to another).[1]

Our application to psychophysical techniques stressed the relative aspects of metacognitive judgments via a paired-comparison ranking methodology (Nelson & Narens, 1980a; other advantages and disadvantages of rankings vs. ratings are discussed by Coombs, 1964). However, we now also use Likert rating scales, with our focus being on the relative aspects of those ratings (by analyzing the rating data via G, as described in Nelson, 1984) as well as on the absolute aspects (via calibration curves), and we do retests on the ratings to assess the stability of the person's threshold; e.g., Nelson et al. (1990; Nelson, McSpadden *et al.*, 1986). Nevertheless, when an investigator is not interested in the absolute aspects of FOK ratings and has the extra time that the ranking procedure usually requires for each subject, the dividend will be, as shown in a large experiment by Lam (1987), that the standard deviation in FOK accuracy across subjects is somewhat smaller for the ranking procedure (SD = .36)

[1]The two kinds of accuracy may yield different conclusions when computed on the same set of data. For instance, confidence may be 100% accurate in the relative sense (i.e., $G = +1.0$) but inaccurate in the absolute sense (e.g., overconfidence, as shown by a calibration curve); this kind of pattern occurred in Nelson, McSpadden *et al.* (1986). In the terminology of Lichtenstein and Fischhoff (1977), relative confidence is reflecting resolution whereas absolute confidence is reflecting calibration, and resolution (in comparison to calibration) "is a more fundamental aspect of probabilistic functioning" (p. 181).

than for the rating procedure (SD = .41), probably because there is a greater tendency for changes in people's thresholds to be neutralized by the ranking procedure. Lam (1987) also showed that the reliability of FOK judgments is greater for rankings than for ratings and that the correlation between FOK rankings and ratings ranges from + .71 to + .87 (the larger correlation is for ratings on a 6-place Likert scale whereas the smaller correlation is for ratings on a 2-place Likert scale). Also, a compromise ranking/rating procedure for use in FOK research has been developed by Shimamura and Squire (1986).

b. Laboratory Paired Associates vs. General-Information Questions. In research where we are interested in the effects of acquisition variables on metamemory, the items are laboratory paired associates such as number—word pairs, whereas in research where we are interested only in the effects of retrieval variables, the items are general-information questions (Nelson & Narens, 1980b; see next paragraph). The former allow for control over the process of acquisition, whereas the latter are fundamentally a version of paired-associate items (e.g., stimulus = "What is the capital of Finland?" and response = "Helsinki"), with the advantages of eliminating the stage of having to teach the items to the person and also having greater stability of recall than does newly learned information.

3. FACTRETRIEVAL Computer Program for Metamemory Research

First, we constructed 300 general-information questions and collected normative data on them (Nelson & Narens, 1980b). Next, we put 240 of those questions into a computer program called FACTRETRIEVAL that tests recall, collects FOK judgments, and tests recognition (Shimamura, Landwehr, & Nelson, 1981). Finally, we enlarged that program into a more sophisticated version called FACTRETRIEVAL2 (Wilkinson & Nelson, 1984) that collects confidence judgments about previous recall, in addition to containing both ranking and rating versions of FOK judgments about upcoming recognition, and that offers many other advantages (e.g., control over the difficulty levels of the items presented to the person, more recognition alternatives per item, assessment of retest reliability of the FOK judgments, and more thorough analysis of the data, including an analysis of response latencies).

B. Some New Findings from Our Research

It is not possible here to summarize all of our metamemory findings from the past 15 years. Instead, we will emphasize those findings that

have not yet appeared in print and will only briefly mention a portion of those already published.

The findings below are organized around several themes, as indicated by the side headings. (All differences mentioned as significant had $p < .05$.)

1. *Amount of Information Deposited in Long-Term Memory Is Important for Metacognitive Monitoring*

a. Our Early Experiments. Our first experiment on metamemory was conducted in 1976 at the University of California, Irvine, and used a paired-comparison ranking methodology on number–word pairs immediately after they had been presented to each subject once during study (a protocol for one subject appears in Nelson & Narens, 1980a). We found that people were very consistent in their FOK judgments, both in terms of transitive FOK paired comparisons (i.e., if Item A is chosen over Item B, and Item B is chosen over Item C, then Item A will have a high probability of being chosen over Item C) and in terms of retest reliability (for near-perfect retest reliability, see Nelson, Leonesio, Landwehr, & Narens, 1986, Fig. 2), but those judgments had nearly no validity for predicting upcoming recognition! This difference between reliability and validity, which we replicated[2] in other unpublished experiments during 1977–1979, was so extreme that for awhile we used a fun-house mirror analogy to describe our subjects' metamemories, wherein what people see when looking in such a mirror is a reliable but nonvalid image of themselves.

After exploring several blind alleys, including the eventually rejected possibilities that (1) people are inherently poor at monitoring their memories, and/or (2) the structure of the underlying items is compromised of both forward and backward traces (in which the person monitored the strength of the forward trace, whereas recognition tapped the strength of the backward trace), we eventually concluded that the lack of FOK validity we had observed was due to the items never having been registered well enough in LTM to be monitored by the metacognitive system. Here is how we came to that conclusion.

b. Effect of Degree of Learning and Retention Interval on FOK Accuracy. In 1979, we conducted an experiment (T. O. Nelson & L. Narens, unpublished) in which three groups (n = 27 or 28 subjects/group) differed

[2]Researchers other than us have also discovered situations in which people have no validity at monitoring their ongoing learning until the items first become recallable after a filled retention interval that exceeds the limits of STM (e.g., Vesonder & Voss, 1985).

in terms of the degree of learning and the delay of the retention test. The retention test consisted of recall, followed by paired-comparison FOK judgments about the person's subjective likelihood of recognizing answers that he or she did not recall, and ended with 4-alternative-forced-choice (4-AFC) recognition test on every nonrecalled item so as to assess the accuracy of the FOK judgments. The results for each group were: (1) the first group, who had an immediate test after one study trial per item, yielded 35% correct recall, and 63% of the subjects had a positive (vs. negative) G for FOK accuracy at predicting the recognition of nonrecalled items (not significant); (2) the second group, who had a delayed test 1 week after acquisition via one correct recall per item, yielded 45% correct recall, and 71% of the subjects had a positive (vs. negative) G (marginally significant); (3) the third group, who had a delayed test 3 weeks after acquisition via one correct recall per item, yielded 25% correct recall, and 86% of the subjects had a positive (vs. negative) G (significant FOK accuracy, $p < .001$). Thus, the level of recall did not determine FOK accuracy (i.e., the first group's level of recall was bracketed by the second and third groups) and the people in all three groups attended to every item during presentation (i.e., all items entered STM), but what mattered was to ensure that the items could be recalled from LTM during acquisition (i.e., only the last two groups showed indications of FOK accuracy). However, from this experiment we could not tell whether the critical factor for FOK accuracy was the degree of learning or the length of the retention interval. Therefore, another experiment was needed to separate the effects of those two factors.

In 1980, we conducted a paired-associate experiment (L. Narens & T. O. Nelson, unpublished) on three more groups: (1) the first group, who had a delayed test 1 week after acquisition via one correct recall per item, yielded 29% correct recall and had little FOK accuracy (mean $G = +.15$); (2) the second group, who had a delayed test 4 weeks after acquisition via one correct recall per item, yielded 15% correct recall and also had little FOK accuracy (mean $G = +.14$); however, (3) the third group, who had a delayed test 4 weeks after acquisition via four correct recalls per item, yielded 38% recall and had significant FOK accuracy (mean $G = +.36$). Thus, not the length of the retention interval but rather the degree of learning is critical for FOK accuracy.

c. Our First Published Experiment on Metamemory. The aforementioned findings led to an experiment in 1981 to establish the importance of the degree of learning on FOK accuracy, which was our first published experiment on metamemory (Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982). We found that 4 weeks after acquisition, FOK accuracy

was nil for items that had originally been learned to a criterion of only one correct recall (median $G = .00$), but FOK accuracy was substantial for items that had originally been overlearned to a criterion of four correct recalls per item (median $G = +.41$). But by what mechanism does the degree of learning affect the accuracy of metacognitive judgments?

d. Mechanism for the Overlearning Effect on Metacognitive Accuracy. The two-part psychological mechanism that seems to underlie this effect is the following:

1. Metacognitive judgments attempt to discriminate between different items and therefore will increase in accuracy as the difference between the to-be-judged items increases, and
2. overlearning may enhance the stability of retention and apparently also increases the differences between items.

For instance, Leonesio and Nelson (1982) found that the degree of learning during acquisition would strongly modulate the gamma correlation between item difficulty (i.e., the number of trials required for the first correct recall during acquisition) and subsequent recall retention: For items that were learned to a criterion of one correct recall during acquisition, this correlation was $+.02$; for items learned to a criterion of two correct recalls during acquisition, the correlation was $-.15$; and for items learned to a criterion of four correct recalls during acquisition, the correlation was $-.25$ (i.e., items that required more trials before the first correct recall were less likely to be remembered during the subsequent retention test).

e. Empirical Support for the Item-Discrimination Mechanism. In accord with the above, Leonesio (1985) found that for overlearned items the correlation between FOK and recognition (namely, $G = +.28$) dropped to nonsignificance when item difficulty was partialled out (namely, $G = +.01$), suggesting that item difficulty—in particular, the differences in difficulty between items—is a factor that modulates the degree of FOK accuracy. Also in accord with Statement (1) above, Nelson, Leonesio et al. (1986) found that FOK accuracy on general-information questions ranged from $G = .00$ for items that are adjacent in the person's FOK rank ordering to $G = +.77$ for discriminating between the top and bottom items in the person's rank ordering. Thus, like a pan balance, people are reasonably fine as measuring devices, but they have limits in terms of the objects between which they can validly discriminate (cf. signal/noise ratio). In this case, the relevant difference is in terms of underlying memorability (analogous to a difference in mass for the pan-balance

case), such that the more different the items, the more likely the metacognitive discriminations are to be valid. This point is important not only for theoretical formulations of metamemory, but also for conclusions about methodology; for instance, a low value of G does not necessarily imply that the task is insensitive or that the person cannot monitor accurately, but rather the obtained value of G reflects a combination of the task, the person's ability to monitor, and the degree of differences among the to-be-monitored items. Using more or less discriminable items will produce greater or lesser degrees of accurate metacognitive discriminations between the to-be-monitored items.

f. Overlearning and the Relation between EOL Judgments, JOL, and FOK Judgments. Because EOL judgments occur prior to acquisition (i.e., before overlearning begins), they cannot tap overlearning but rather can only tap item difficulty.[3] However, in contrast to EOL judgments, JOL can tap both item difficulty and the degree of learning (because JOL occur after acquisition). Not surprisingly, therefore, subsequent recall retention is predicted significantly better by JOL ($G = +.31$) than by EOL ($G = +.12$), as shown by Leonesio and Nelson (1990). Also during the retention session, the recognition of nonrecalled items is predicted as well by JOL (which had been made 4 weeks earlier) as by FOK judgments (which had been made immediately prior to the recognition test), and those two kinds of judgments are not themselves highly correlated with each other and therefore may tap different aspects of memory (see Leonesio & Nelson, 1990).

2. *FOK May Be Perfectly Valid at Tapping a Large Number of Aspects of LTM but the Accuracy of FOK for Predicting Criterion Performance May Nevertheless Be Imperfect*

There are at least two possible reasons, in addition to the methodological one mentioned above (i.e., items not different enough for the person to be able to discriminate between them), that the observed FOK accuracy may underestimate the actual FOK accuracy at monitoring information in LTM.

[3]Moreover, Leonesio and Nelson (1990) showed that those EOL judgments have far-from-perfect accuracy at monitoring item difficulty (e.g., the mean correlation between EOL judgments and the number of trials required to learn the various items in a constant-study-time situation is only $G = -.22$), and the relatively low magnitude of this correlation is not due to inadequate range in the number of trials required to learn the various items (e.g., the mean correlation between the number of learning trials and subsequent recall 4 weeks later was $G = -.48$).

1. No single criterion task may tap the full set of information tapped by the FOK. As just one example of this possibility, consider the typical criterion task—namely, recognition—that is used to validate the accuracy of FOK. We know that recognition does not completely tap the information in memory (e.g., savings occurs for nonrecognized items, Nelson, 1978), so some of the information in memory that is tapped by FOK may be overlooked by recognition (and perhaps vice versa, of course). Given the view that memory is multidimensional rather than unidimensional (e.g., Bower, 1967), it is even possible that the FOK may be tapping more aspects of memory than any single criterion task. That is, different criterion tasks tap different aspects of memory, and the current view is that no particular task is strictly more sensitive (in the technical sense; Nelson, 1978) than all other criterion tasks (for a review of the rapidly growing literature that shows how different tasks are dissociated from one another and tap different aspects of memory, see Richard-Klavehn & Bjork, 1988).

2. FOK does not detect small amounts of new information coming into memory that may affect criterion performance. We recently discovered a situation in which the FOK can be less sensitive than recall for detecting information in memory (Jameson, Narens, Goldfarb, & Nelson, 1990). In that research, a nonrecalled general-information answer was very briefly flashed while the person was attending to the corresponding general-information question. The very brief flash contained either the correct answer or a nonsensical answer. Following the flash, the person either (1) immediately attempted to recall the answer to the question and then immediately gave an FOK judgment about whether he or she knew the answer (Experiment 1), or (2) immediately gave an FOK judgment without any intervening attempt at recall (Experiment 2, which was run as a control in case the effects of the flash dissipated during the immediate-recall phase in Experiment 1, before the FOK judgment occurred).

The results, summarized in Table I, show that the new information added to memory by the very brief flash affected recall without affecting FOK. This is in accord with the aforementioned findings that the FOK can tap only those aspects of information that previously had been well-established in LTM and does not detect new incoming information. Consistent with such a conclusion, the FOK can validly discriminate between nonrecalled items that will soon have the correct answer flashed tachistoscopically (Nelson *et al.*, 1984, Exp. 1). Taken together, these results from our 1984 and 1990 research suggest that the residual information in LTM that is tapped by the FOK can be augmented by incoming flashed information that the FOK does not detect. Whether this incoming flashed

TABLE I

EFFECT OF A PERCEPTUAL FLASH (OF THE CORRECT ANSWER VS. NONSENSE) ON THE SUBSEQUENT PROBABILITY OF RECALL AND THE FEELING OF KNOWING (FOK)[a]

	Answer that was flashed	
Dependent variable	Correct answer	Nonsense
Experiment 1: *p*(recall)	.28	.10
Experiment 1: FOK rating	5.4	5.2
Experiment 2: FOK rating	6.1	6.1

[a]The entry for *p*(recall) is the mean (across subjects) of each individual subject's *p*(recall). The entry for FOK rating is the mean (across subjects) of each individual subject's median FOK rating (higher values indicate a stronger feeling of knowing). Although flashing the correct answer (vs. nonsense) yielded a significant improvement in recall beyond the reminiscence that occurred in the nonsense condition, no significant effect occurred on the feeling of knowing. Data are from Jameson *et al.* (1990).

information should be conceptualized as residing in STM or in unconscious memory (cf. Marcel, 1983) is an open question, whose answer may have ramifications for conceptions about the limits of metacognitive monitoring.

Thus, the FOK can tap LTM information that by itself is insufficient to trigger correct recall, whereas recall can be based on the conglomerate of both the preflash information in LTM and a boost from flashed information that the FOK does not tap. From this research, we now know that at least some information in the overall memory system is not tapped by FOK, and therefore the question arises concerning the degree to which people do have direct (or privileged) access to their own idiosyncratic memories.

3. *Privileged Access*

Do people have privileged access to idiosyncratic information in their memories about the to-be-retrieved items? We examined this question in two ways.

a. Judge/Observer Experiments. We (T. Jameson, T. O. Nelson, R. J. Leonesio, & L. Narens, unpublished) modified our standard FACTRETRIEVAL paradigm as follows. One person (designated the Target sub-

ject—the standard subject in FACTRETRIEVAL) went through recall until missing the answers to 15 questions. Then he or she made FOK rankings of those 15 items. Meanwhile—and here's the new twist—while the Target was going through recall, another person (designated the Observer) observed the Target's performance during recall (i.e., the Observer saw the Target's face, saw how long the Target paused to think about the answer to each question, saw how well the Target did on related questions, and saw what the Target typed as a recall response to a given question). Then the Observer went to another computer room and independently ranked those same 15 items in terms of how likely the Target would be to recognize the correct answer to each missed item. Finally, yet another person (designated the Judge), who never saw the Target or the Target's answers during recall, also ranked those same 15 items in terms of how likely the Target would be to recognize the correct answer to each missed item. Subsequently the Target went through a 4-AFC recognition test on each item, so as to provide the criterion performance that allowed us to assess the predictive accuracy of the Target's, Observer's, and Judge's predictions about the Target.

The hypothesis we tested was the following. The Judge would have some above-chance accuracy at predicting the Target's recognition performance, based on the Judge's knowledge of the general difficulty of each of the various items. The Observer would have the same knowledge about general item difficulty that the Judge had, but also by virtue of having watched the Target during attempted recall of each item would have some specific extra knowledge about what the Target might know (e.g., if the Target paused to think awhile before answering or made a close guess at the answer), and therefore the Observer would be more accurate than the Judge at predicting the Target's subsequent recognition. The Target would have all of the above information, plus "privileged" information about his own idiosyncratic memory (e.g., remembering that he or she had learned a particular item in high school) and therefore should be the best possible predictor of his or her subsequent recognition performance.

The results, shown in the left side of Fig. 7, generally confirmed the aforementioned hypothesis about the relative predictive accuracy of the Target, Observer, and Judge for predicting the Target's subsequent recognition performance. All three sets of predictions had above-chance accuracy, and the Target's predictive accuracy was significantly greater than the Judge's, with the Observer's predictive accuracy being intermediate between them (but not significantly different from either the Target or the Judge).

Because the overall predictive accuracy in that experiment was somewhat low, we ran another experiment containing a few methodological

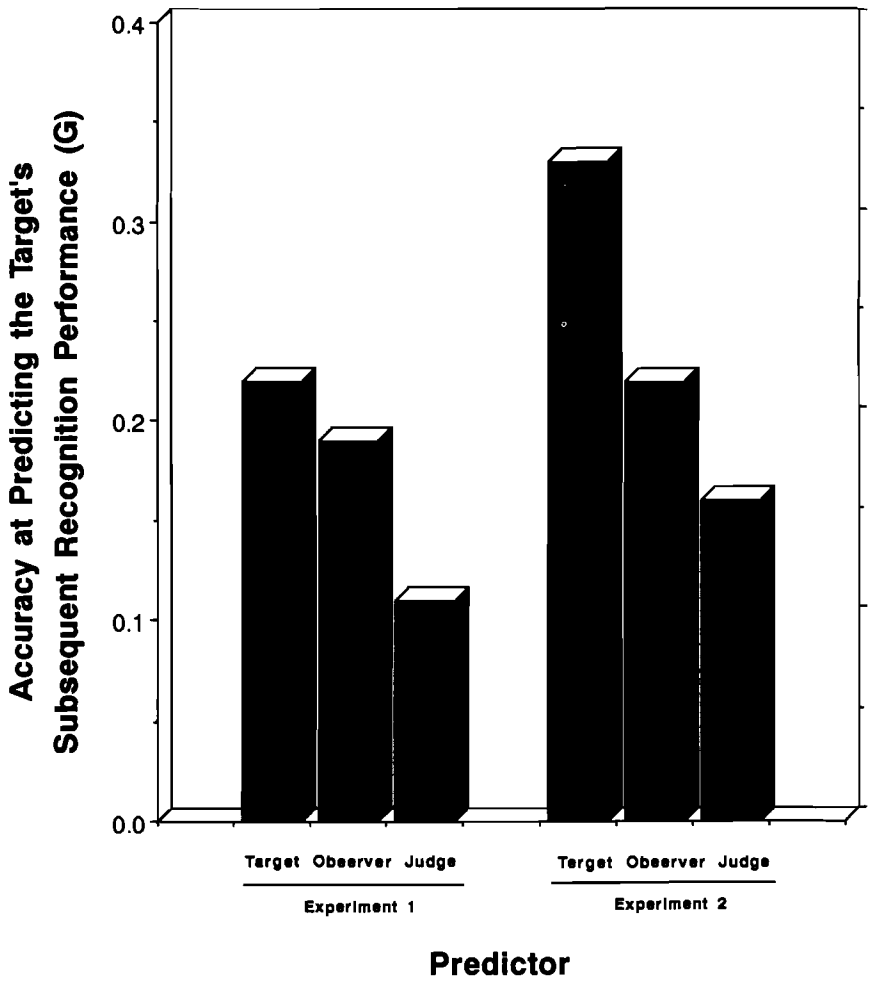

Fig. 7. Accuracy at predicting the target's subsequent recognition performance on non-recalled items (in terms of mean *G*) for three predictors (the Target, the Observer, and the Judge) in two experiments.

changes. In particular, during recall the items were not randomly sampled but instead were sampled more systematically to utilize the full range of general-information questions in FACTRETRIEVAL, and the recognition test was made less noisy by using 8-AFC instead of 4-AFC (both of these changes yield greater overall predictive accuracy; see above). Also, instead of being in the same room with the Target, the Observer watched

the Target through a one-way mirror from an adjacent room during the recall phase.

The results, reported in the right side of Fig. 7, showed greater predictive accuracy overall than in the previous experiment. Most important, the general pattern and the qualitative conclusions from the statistical tests were the same as in the previous experiment. Thus people apparently do have idiosyncratic information at their disposal during retrieval, and this idiosyncratic information can benefit predictions about their subsequent memory performance. But does this idiosyncratic information about their own memories yield the best possible predictive accuracy about their subsequent performance, or is there a way to improve predictive accuracy even more?

b. Normative Predictions vs. the Individual's Own FOK Predictions. It is worth mentioning that Nelson, Leonesio *et al.* (1986) found that for predicting an individual's subsequent memory performance on currently nonrecalled items, the individual's own FOK predictions were significantly better (mean $G = +.28$) than normative FOK predictions (means $G = +.12$) derived from the average of his or her peers' predictions about their own memory performance, but the individual's FOK predictions were significantly worse than predictions derived from the normative probability of correct recall (mean $G = +.38$). We made several attempts to induce individuals to utilize (while making FOK judgments) their estimates of normative recall, in hopes that this might yield an improvement in FOK accuracy. Unfortunately all of our attempts failed to improve people's FOK accuracy, perhaps because people are poor at trying to intuit the normative probability of recall on items that they themselves cannot recall (Nickerson, Baddeley, & Freeman, 1987). However, M. Calogero (unpublished research conducted in our laboratory) found that people who are given the normative probability of recall as they make FOK judgments for each item do have significantly greater FOK accuracy ($G = +.58$) than other people who are not given those normative probabilities ($G = +.40$), indicating that people will utilize normative information when it is available (in his experiment, the accuracy from predictions derived solely from the normative probability of recall was $G = +.55$). Next, we turn to the question of what the information is that does underlie people's FOK judgments.

4. *Some Factors Underlying People's FOK Judgments (versus FOK Accuracy)*

To inquire about whether a given factor "affects FOK accuracy" is to ask whether the factor affects the relationship between the FOK judg-

ments and the criterion task (e.g., in the above-mentioned research this was the relationship—as assessed by *G*—between FOK judgments and subsequent recognition performance). By contrast, to inquire about whether a given factor "affects FOK judgments" is to ask whether the factor affects the magnitude of FOK, as assessed by the median FOK rank or the median FOK rating. These two possible meanings of "an effect on FOK" are mathematically independent of each other. The former kind of effect was examined above. The latter kind is examined next.

a. Overlearning Affects Not Only FOK Accuracy but also Affects the Magnitude of FOK. Nelson *et al.* (1982) reported that the median FOK rank varied across items that differed in the degree of original learning: Items originally learned to a criterion of one recall per item had a median FOK rank of 5.8; items with one additional overlearning trial had a median FOK rank of 6.8; and items with three additional overlearning trials had a median FOK rank of 8.4.

This effect of overlearning on the magnitude of metacognitive judgments was extended recently by Leonesio and Nelson (1990). They investigated a situation in which people (1) made JOL at the end of acquisition, and (2) subsequently made FOK judgments 4 weeks later on items incorrectly recalled during the retention test (using a retention-session procedure similar to the one from Nelson *et al.*, 1982). A major finding, shown in Fig. 8, was that overlearning has a greater effect on JOL than on subsequent FOK judgments.

In all of the aforementioned experiments, the overlearning trials were a combination of study-test trials (e.g., "three additional overlearning trials" meant three additional overlearning study trials and three additional overlearning test trials). But which portion—overlearning study or overlearning test—were the FOK judgments being affected by?

b. Overlearning Study Trials vs. Overlearning Test Trials. An experiment by T. O. Nelson, T. Rideout, and R. J. Leonesio (unpublished) had people learn a paired-associate list in which one-third of the items were learned to a criterion of one correct recall per item, another one-third had six overlearning study trials after the item was correctly recalled, and the remaining one-third had six overlearning test trials after the item was first correctly recalled. Four weeks later, the median FOK rank for items not recalled on the retention test was 6 for the items that had originally been learned to a criterion of one correct recall, 8 for the items that had received six overlearning study trials, and 8 for the items that had received six overlearning test trials (each of the latter two sets of items differed significantly from the first but did not differ from each other). Thus, both the overlearning study trials and the overlearning test

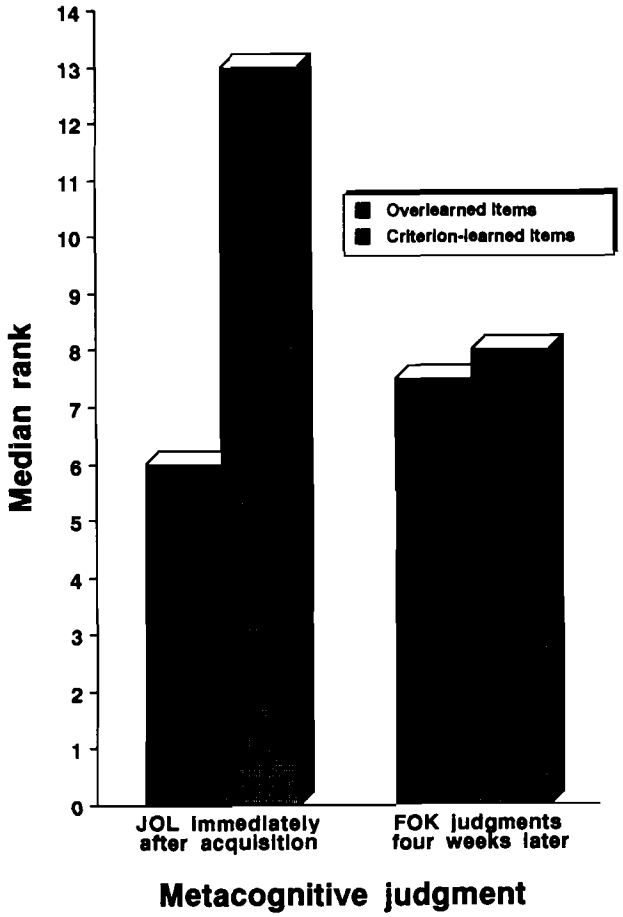

Fig. 8. Median JOL rank and FOK rank (1 = low) for items that were acquired to a criterion of one correct recall (leftmost in each pair of bars) vs. four correct recalls. The effect of overlearning is greater on JOL than on FOK.

trials affect the magnitude of subsequent FOK, and to approximately the same degree.

c. Actual Overlearning vs. Claimed Overlearning. Although overlearning has an effect on FOK, we wondered if that effect was a direct one or instead was mediated by whether the person was aware that the item had been overlearned. An experiment by T. O. Nelson and S. Gilispie (unpublished) attempted to tease those two factors apart.

The college-student subjects each received study-test trials on a paired-

TABLE II

EFFECT OF ACTUAL FREQUENCY OF PREVIOUS RECALLS VS. CLAIMED FREQUENCY OF PREVIOUS RECALLS ON SUBSEQUENT FOK JUDGMENTS[a]

Claimed frequency of previous recalls	Actual frequency of previous recalls		
	1	5	Overall
1	10	**10**	10
5	**16**	18	17
Overall	12	14	

[a]The entry is the median (across subjects) of each individual subject's median FOK rank for the items in that cell (higher values indicate a stronger feeling of knowing). Although both the actual and the claimed frequency of previous recalls had an overall significant effect, notice that the feeling of knowing is affected more by the person's claimed frequency than by the actual frequency of previous recalls. This can be seen in two ways: by comparing the effect manifest in the row marginals with the effect manifest in the column marginals, or, perhaps even better, by comparing the two internal-cell values shown in boldface (i.e, the person's feeling of knowing was stronger for items that had actually been recalled once but which he or she believed to have been recalled five times than for items that were believed to have been recalled once but that had actually been recalled five times).

associate list, wherein 16 items were learned to a criterion of one correct recall per item while the remaining 16 were overlearned (five correct recalls per item). A retention session occurred 3–7 weeks later (this difference in retention interval had no effect on recall or on FOK judgments and therefore will not be discussed further), consisting of three stages: (1) recall of every item, followed by (2) several judgments on every nonrecalled item, the two most pertinent for present purposes being the students' forced-choice frequency judgments of whether a given item had originally been learned to a criterion of one or five correct recalls and the students' FOK judgments (which occurred either before or after—counterbalanced—the other judgments had been made), followed by (3) 8-AFC recognition.

Table II shows that the person's FOK was related more to his or her

TABLE III

EFFECT OF ACTUAL FREQUENCY OF PREVIOUS RECALLS VS. CLAIMED FREQUENCY OF PREVIOUS RECALLS ON THE SUBSEQUENT RECOGNITION THAT THE FOK IS ATTEMPTING TO PREDICT[a]

Claimed frequency of previous recalls	Actual frequency of previous recalls		
	1	5	Overall
1	34	42	38
5	41	53	47
Overall	38	47	

[a]The entry is the mean (across subjects) of each individual subject's percentage correct recognition for the items in that cell. Both the actual and the claimed frequency of previous recalls have an overall significant effect (and to the same degree).

claimed frequency of previous recalls than to his or her actual frequency of previous recalls. Although this can be seen by comparing the two pairs of marginals, it is perhaps most evident by noticing that the FOK was greater for items that the person believed he or she had previously recalled five times but actually had been recalled only once (median FOK = 16) than for items believed to have been previously recalled only once but that had actually been recalled five times (median FOK = 10). Put another way, most of the effect of overlearning on FOK judgments is mediated by the person's beliefs about whether the items had been overlearned; i.e., the person's memory of prior overlearning (presumably taken together with rules of inference) mediates the effect of prior overlearning on FOK judgments. This is quite different from a direct or "automatic" effect of overlearning, in which the effect on a particular dependent variable occurs regardless of whether or not the person remembers that one item had occurred more frequently than another item during study.

However, this pattern of FOK judgments is not entirely reflecting the recognition that the person was attempting to predict, as shown in Table III. In particular, the percentage correct recognition was affected approximately equally by both the claimed frequency (probably due to other differences in the items—the row headings are labels for the subjects' aggregations, not for an independent variable) and the actual frequency.

This greater effect on the FOK by the claimed frequency than by the actual frequency can also be seen in another way. We correlated (across items for each subject) the FOK with three other variables: (1) claimed frequency of previous recalls, (2) actual frequency of previous recalls, and (3) recognition performance. The mean correlations were, respectively, + .40, + .23, and + .16. Although all three correlations are significantly greater than zero, FOK is more related to the claimed frequency of previous recall than to the actual frequency of previous recall; moreover, FOK is more related to the claimed frequency of previous recall than to the recognition that the FOK is attempting to predict.

These results suggest a new hypothesis—which we dub the No-Magic Hypothesis—for how FOK judgments should be conceptualized, namely, the person (1) considers particular recallable properties of the to-be-retrieved item (e.g., "I recalled it few/many times on previous occasions"), in conjunction with (2) rules about how those properties are related to the subsequent criterion performance that the person is trying to predict (e.g., "subsequent recognition is more likely for an item that I previously recalled many times than for an item that I previously recalled few times"). Notice that this way of making FOK judgments would utilize only suprathreshold information about remembered attributes of the item (including incorrectly remembered suprathreshold information!), along with rules for how to utilize that information in the FOK judgments. Thus, according to the No-Magic Hypothesis, the FOK does not reflect any monitoring of unconscious information at all. Put another way, the FOK does not directly monitor a given unrecalled item in memory, but rather the FOK monitors recallable aspects related to that item, such as the item's acquisition history or partial/related recalled components.

5. *Learning-to-Learn Effects for FOK Judgments*

Can FOK accuracy be improved by sheer practice? This question was explored in two experiments by T. O. Nelson, R. J. Leonesio, and L. Narens (unpublished) that were modifications of the standard FACTRETRIEVAL computer program. In Experiment 1, each subject attempted to recall answers to general-information questions until 45 questions had been missed. Then the subject went through three blocks of 15 nonrecalled items per block. For the first block, the subject made FOK judgments and then received a 7-AFC recognition test on each of the 15 items, followed by the same sequence for the second block and then for the third block. After each block, 58 subjects received feedback (seeing a table of their FOK judgments and their recognition performance on every item), and another 58 subjects received no feedback.

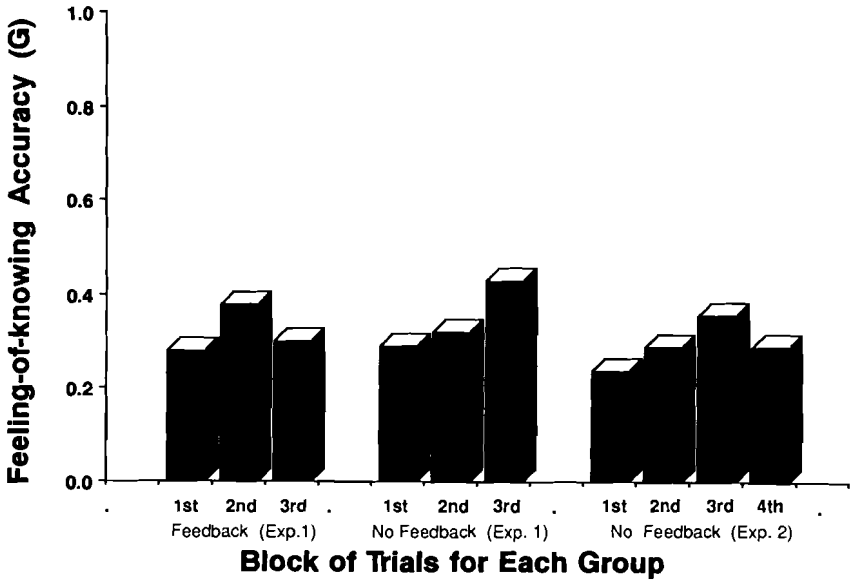

Fig. 9. Feeling-of-knowing accuracy (in terms of mean G) on each block of trials for each group (Feedback or No Feedback) in two experiments.

The resulting FOK accuracy is shown in the left and middle portions of Fig. 9. There was no significant change in the feedback group and there was not even a trend of uniformly increasing FOK accuracy (left side of Fig. 9). However, in the no-feedback group a small but significant increase in FOK accuracy occurred from Block 1 to Block 3, $t(57) = 2.69$ (middle portion of Fig. 9).

In an attempt to replicate and extend this finding that by sheer practice without feedback, people can learn to make FOK judgments more accurately, data were examined in Experiment 2, where only the no-feedback condition was run. Following recall on general-information questions so as to isolate 40 nonrecalled items, each of 36 subjects went through four blocks of 10 items per block. During each block, the subject made FOK judgments, followed by a 4-AFC recognition test on each item, and no feedback was given between blocks.

The resulting FOK accuracy, shown in the right side of Fig. 9, was not significantly different across the four blocks and the pattern was not even in accord with the hypothesis of uniformly increasing FOK accuracy across blocks. Thus we conclude that there is no systematic change in FOK accuracy across blocks of items, and the effect of learning to learn

by sheer practice on FOK judgments is small if it occurs at all (i.e., negligible).

Looked at differently, the FOK accuracy of sizeable groups of people appears to be fairly stable across blocks of FOK judgments and does not change in any systematic way. This stability of a group's FOK accuracy can be a methodological advantage because it allows the use of within-subject designs to assess the effect of a given independent variable on FOK accuracy. At the same time, however, researchers should be aware that an individual subject's scores of FOK accuracy vary widely for all known FOK procedures and items (Nelson, 1988), such that (by present-day methodology) individual differences in FOK accuracy are unstable. An important challenge for future research is to discover ways of achieving stable individual scores of FOK accuracy, for use both in theoretical-memory research (e.g., so that such scores can be correlated meaningfully with other individual differences) and in applied settings (e.g., diagnostic classification of a newly admitted hospital patient or a poorly performing student). For instance, perhaps a standardized set of items can be found that yields stable performance (as in an optometrist's eye test) or perhaps a very large number of items would have to be examined for each individual (as in traditional psychophysics experiments). Whether or not these conclusions about stable FOK accuracy for groups of subjects, in conjunction with unstable individual differences in FOK accuracy, extend to other kinds of metamemory judgments (e.g., JOL) is an open question.

6. *Relation between Metacognitive Monitoring and Metacognitive Control during Acquisition: The Allocation of Self-Paced Study Time*

In the theoretical framework above, Fig. 4 shows some interactions during acquisition between metacognitive monitoring processes and metacognitive control processes. We have conducted several experiments pertaining to that topic.

Nelson and Leonesio (1988) found that the allocation of study time is not simply a direct effect of item difficulty but rather is mediated by the person's EOL and FOK judgments about item difficulty (similar to the person's memory of overlearning being a mediator between the actual amount of overlearning and subsequent FOK, as described earlier). Thus the person's metacognitive monitoring may mediate much of the effect of a given independent variable on the person's control of cognitive processing. This illustrates one way in which we can sometimes improve our predictions about how people will control their own cognitive processing if we obtain metacognitive monitoring judgments (which comprise part of

the input people use when they adjust the parameters of their control processes). However, we also want to know the conditions under which people's metacognitive monitoring is to some degree irrelevant to their control processes during acquisition. A few specific examples will help to illustrate.

In Nelson and Leonesio's Experiment 1, the mean correlation between EOL and study time was −.3 (i.e., items believed to be harder were allocated more self-paced study time). However, there was incomplete compensation for differences in item difficulty during self-paced study. That is, the mean correlation between EOL and recall after self-paced study was not zero (which would have indicated that the extra study time allocated to the items believed to be harder had completely compensated for the differences in item difficulty), but rather was a hefty $G = +.48$. Thus, items perceived as easier were more likely (than items perceived as harder) to be recalled, even after self-paced study in which some extra study time had been allocated to the items perceived as harder. In Experiments 2 and 3, Nelson and Leonesio (1988) found that this incomplete compensation also occurs for FOK judgments and the subsequent self-paced allocation of study time, and the absolute magnitude of the correlation between FOK and self-paced study time was not very large.

Therefore, in a follow-up study (T. O. Nelson & R. J. Leonesio, unpublished), we decided that the computer rather than the person would control the study time per item (cf. Groen & Atkinson, 1966), but the computer would base the distribution of study times on the person's FOK judgments. Four different ways of allocating various study times were examined. All four groups ($n = 52$ Ss/group) made FOK judgments on nonrecalled general-information items (as in Nelson & Leonesio, 1988, Experiment 3), after which the groups received different distributions of study time per nonrecalled item (but the same total study time for the entire set of nonrecalled items): (1) The constant-time group had 4.3 sec of study on every item; (2) the random-assignment-of-different-times group had a random assignment of 8 vs. .5 sec per item, regardless of FOK[4]; (3) the More-Time-To-Low-FOK group had 8 sec per item on the 50% of the items with the lowest FOK and .5 sec per item on the 50% with the highest FOK; and (4) the more-to-high-FOK group had 8 sec per item on the 50% with the highest FOK and .5 sec per item on the 50%

[4]A prerequisite for interpreting the results was that the difference in study time was substantial enough to produce differences in subsequent recall, and this prerequisite was confirmed when the mean correlation between the amount of study time and subsequent recall in the random-assignment-of-study-time group was significantly greater than zero ($G = +.39$).

with the lowest FOK. Subsequently, the mean percentage correct recall for each of the four groups was 63.9, 64.5, 64.4, and 56.7, respectively. The more-time-to-high-FOK group did significantly worse than the other three groups, which did not differ significantly. These results demonstrate that allocating extra study time to items that people believe they already know (but currently cannot recall) is an inefficient allocation of extra study time, but the surprising finding is that allocating extra study time to items which people believe they don't already know was no better than allocating the same amount of time to all items and wasn't even better than allocating different amounts of study time randomly! Perhaps with a wider range in the allocation of study time per item (e.g., as would be shown by a higher value—ideally, near +1.0—of the correlation in footnote 4) there would be a greater effect of allocating extra study time to items which are believed to be more difficult. Or perhaps the correlation between EOL/FOK and actual item difficulty is too low (i.e., perhaps inaccurate monitoring of item difficulty is giving misleading information for the control of the allocation of study time).

7. *Relation between Metacognitive Monitoring and Metacognitive Control during Retrieval: Termination of Memory Searching*

In the theoretical framework, interactions between metacognitive monitoring processes and metacognitive control processes also occur during retrieval, and we have recently begun several lines of research on this topic.

a. Role of "Preliminary FOK Judgments" Prior to Searching for an Answer. The first construct in the hypothesized retrieval process (see Fig. 5) is a Preliminary FOK Judgment in which the person decides whether to terminate the retrieval stage or to proceed with a search of memory for the to-be-retrieved item. Empirical confirmation of this hypothetical construct requires that the latency of such FOK judgments be shorter than the latency of correct recall for the requested item. Previous research by Reder (1987, 1988) had confirmed this speculation by asking people either to make binary (i.e., do/don't know) FOK judgments or to retrieve the sought-after answer, and she found that the FOK latencies were indeed shorter than the latencies for retrieving the correct answer.

We utilized Reder's finding to explore a related question, namely, should we conceptualize the Preliminary FOK as consisting of a single FOK component that taps only the presence of information in memory (which we refer to as the *single-counter FOK hypothesis*) or should the conceptualization be in terms of two FOK components, one of which taps the presence of information in memory and the other of which taps the

absence of such information (which we refer to as the *dual-counter FOK hypothesis*)? That is, for the single-counter FOK hypothesis, imagine that there is one counter which is incremented as information comes into the metacognitive system to indicate the presence of the item in memory (e.g., memories of having recalled the item during a recent acquisition session, etc.), and assume that when the amount of accumulating information exceeds a threshold, the FOK response of "will recognize" occurs. Then the only way that a "won't recognize" FOK judgment would occur is by default (i.e., time passes without enough information accumulating to indicate that the item is present in memory and eventually the person responds with "won't recognize").

By contrast, for the dual-counter FOK hypothesis, imagine that one counter—the Affirmative-FOK counter—keeps track of accumulating affirmative information that the item is stored in memory, while another counter—the Negative-FOK counter—keeps track of accumulating negative information that the item is not stored in memory. Thus the "won't recognize" judgment does not occur by default but rather occurs by an accumulation of information directly supporting that judgment. Although the threshold mechanism for the dual-counter FOK hypothesis is unspecified, one possibility is that a difference threshold must be exceeded (i.e., the absolute magnitude of the difference between the value of the Affirmative-FOK counter minus the value of the Negative-FOK counter must exceed a threshold; then if the difference is positive, the "will recognize" response occurs, whereas if the difference is negative, the "won't recognize" response occurs, and if the value does not yet exceed the threshold, then the search for more positive/negative information continues).

We (T. Schreiber, T. O. Nelson, & L. Narens, unpublished) tested the single-counter vs. dual-counter hypotheses by having each person (n = 10) say aloud as quickly as possible an FOK judgment from a 6-place Likert scale (1 = "completely certain I would not even recognize the correct answer") upon presentation of each of 239 general-information questions from the FACTRETRIEVAL program. The single-counter hypothesis predicts that the latencies of making FOK judgments should be an inverse monotonic function of the FOK judgment; e.g., a judgment of FOK = 6 should have the shortest latency, and a judgment of FOK = 1 (the default judgment if no threshold for accumulating affirmative information is eventually exceeded) should have the longest latency. By contrast, the dual-counter hypothesis predicts that the latencies of making FOK judgments should be a nonmonotonic function of the FOK judgments; i.e., judgments of FOK = 6 and FOK = 1 should have the shortest latencies, with the FOK judgments nearer the center of the FOK rating scale having the longest latencies.

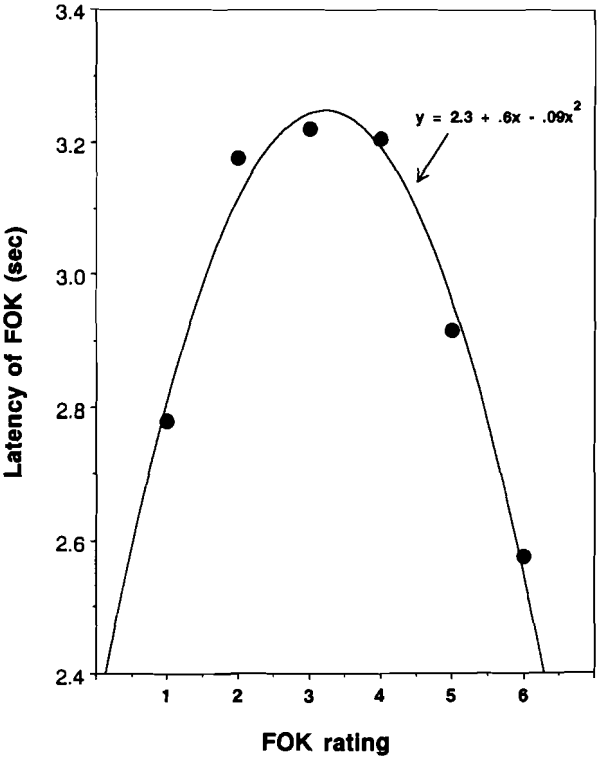

Fig. 10. Median of individual-subjects' median latency of FOK ratings is shown for each FOK rating.

Figure 10 shows the results, in terms of medians of individual-subject median latencies for each FOK rating category.[5] A nonmonotonic pattern is obvious in Fig. 10, which confirms the dual-counter hypothesis that there is both an Affirmative FOK to tap the presence of information in memory and a Negative FOK (cf. "knowing not" in Kolers & Palef, 1976) to tap the absence of information in memory and to allow for quick termination of retrieval when an answer is not known (as in the right branch of Fig. 5).

[5]After a given subject finished making a Preliminary FOK judgment on the 239th item, a recall test occurred on every item to assess the recall latencies. At every FOK rating, the latency of correct recall was longer than the corresponding latency of FOK judgment (the median of individual-subject median latencies of correct recall ranged from a low of 2.9 sec for FOK = 6 to a high of 5.6 sec for FOK = 1), replicating and extending the findings from Reder (1987).

This section focused on how a rapid Preliminary FOK judgment can affect the latency of the termination of retrieval. Next we consider how the ongoing FOK (see the center of Fig. 5) might affect the latency of retrieval termination for longer-latency omission errors[6] (right branch of Fig. 5).

b. Relation between Ongoing FOK and the Latency of Recall Errors. Our previous research had shown that the correlation between FOK and the latency of omission errors is substantial (Nelson *et al.*, 1984) and this was replicated by Nelson *et al.* (1990), where that correlation was $G = +.60$. This is in accord with the center and right branches of Fig. 5, in which the person's FOK affects whether the overall retrieval stage will be continued or terminated, namely, people terminate the retrieval stage more quickly on items for which they have a low FOK. Given this substantial correlation, if FOK accuracy could somehow be improved, so might efficiency improve for terminating retrieval (i.e., people might be more likely to continue to search for items that are known and be more likely to terminate the retrieval stage for items that are not known). In accord with this idea, neuropsychological patients such as Korsakoffs have poor FOK accuracy (Shimamura & Squire, 1986) and they also tend to terminate prematurely when searching memory and therefore have poor recall performance (Hirst, 1982, p. 454).

Contrary to the aforementioned substantial correlation between FOK and the latency of omission errors, the correlation between FOK and the latency of commission errors is nil for general-information items (Nelson *et al.*, 1984; also replicated by Nelson *et al.*, 1990, where that correlation was +.03) or even slightly negative for laboratory paired associates (Nelson *et al.*, 1982). Thus only a negligible correlation occurs between FOK judgments and the amount of time that the person searches before outputing a commission error. Such a conclusion is not in conflict with the conclusion about the relation between FOK and omission errors because the mechanisms are presumably different for omission errors (right branch of Fig. 5) vs. commission errors (left branch of Fig. 5). The latter involves other metacognitive judgments besides FOK, and we consider those other metacognitive judgments next.

c. Relation between Metacognitive Confidence Judgments and the Latency of Recall. Nelson *et al.* (1990) asked people to make retrospec-

[6]For most retrieval situations, the latency of omission errors is substantially longer than the latency of correct responses (e.g., in Nelson, *et al.*, 1984, the average correct-response latencies ranged from 9.8 to 15.2 sec across three groups of subjects, whereas the average omission-error latencies ranged from 20.5 to 25.1 sec. In Nelson *et al.*, 1990, where the responses were spoken rather than typed into a computer, the average correct-response latency was 2.9 sec, whereas the average omission-error latency was 9.9 sec.).

tive confidence judgments about the accuracy of their previous recall response immediately after outputing that response. In contrast to the aforementioned nil relation between FOK and the latency of commission errors ($G = +.03$), the correlation between retrospective confidence judgments and the latency of commission errors was substantial ($G = -.40$). The direction of this correlation indicates that people have greater confidence for items that they retrieve quickly. This different pattern for FOK vs. confidence judgments (when computed on the identical set of commission-error latencies) confirms the idea that different metacognitive judgments tap different aspects of memory (additional evidence is reported in Leonesio & Nelson, 1990) and is in accord with the left branch of Fig. 5, where the final component that affects the output of commission-error responses is the person's confidence. Also in accord with that idea, the correlation between retrospective confidence judgments and the latency of correct recall is substantial ($G = -.55$) (Nelson *et al.*, 1990). Thus people have greater confidence for answers that they retrieve quickly, regardless of whether those answers are wrong (i.e., commission errors) or correct.

Given the aforementioned facts that (1) confidence judgments are correlated with the latency of recall, regardless of whether that recall is correct or a commission error (Nelson *et al.*, 1990), and (2) latencies tend to be longer for commission errors than for correct recall (Nelson *et al.*, 1984; replicated in Reder, 1987, and in Nelson *et al.*, 1990) the following question arises: Is anything other than the latency of recall responsible for retrospective confidence being greater for correctly recalled items than for commission errors? That is, perhaps the only reason that confidence is greater for correctly recalled answers than for commission errors is because the former are recalled faster than the latter, with confidence being based entirely on recall latency.

A test of this Confidence-Determined-Entirely-by-Latency Hypothesis occurred in an experiment by T. O. Nelson and M. Calogero (unpublished) in which people went through the recall phase of FACTRETRIEVAL2 on 106 general-information questions and made retrospective confidence judgments immediately after outputing each answer. For data analysis, the items were categorized into correct recalls and commission errors (omission errors are irrelevant because retrospective confidence judgments did not occur after omission errors). For each subject, the items in each of those two categories were Vincentized into six blocks (i.e., the first block consisted of the one-sixth of the items that had the shortest recall latencies, and so on to the sixth block, consisting of the one-sixth of the items for which that subject had the longest recall latencies), and the median response latency of the items within each block was

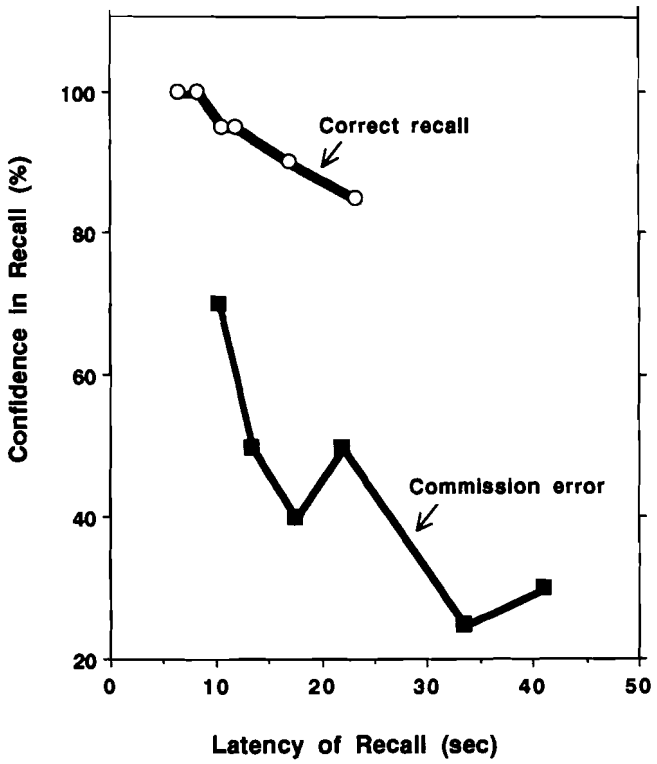

Fig. 11. Confidence in recall for each category of recalled answer (correct recall or commission error) at each Vincentized block of one-sixth of the items ordered in terms of the latency of recall. The value on the abscissa corresponds to the mean of the individual subjects' median latencies, and the value on the ordinate corresponds to the mean of the individual subjects' median confidence for that sixth of the items.

determined. Then the median confidence judgment was determined for the items in each of those six blocks. Finally, the means (across subjects) of those individual-subject median latencies and median confidence judgments were plotted in bivariate form, separately for the categories of correct recall vs. commission error (see Fig. 11).

Several conclusions can be drawn from Fig. 11. First, as expected from the correlations mentioned earlier, retrospective confidence tends to decrease as the latency of recall increases, both for correct recalls and for commission errors. Second and alternatively, the greater the confidence, the shorter the recall latency tends to be (cf. left branch of Fig. 5). Third and most important for the Confidence-Determined-Entirely-by-Latency Hypothesis, the two curves clearly do not lie one atop the other (espe-

cially notice the portions of the curves for latencies in the range from 10 to 22 sec), whereas one curve atop the other should have occurred if the latency of recall had been the sole determiner of confidence. Thus, something in addition to the latency of recall is affecting the person's confidence and is producing greater confidence for correct recall than for commission errors. We do not yet know what that something is, but future research should attempt to isolate and identify it.

IV. Applications of Our Methodology to Other Areas

Our methodology, including versions of the FACTRETRIEVAL paradigm, has been used in many other laboratories to investigate both the domain of traditional memory research and related areas. We and our past and present co-workers have also applied this methodology to the following areas. Developmental psychology: Butterfield, Nelson, and Peck (1988) examined changes in metamemory across the life span from 6 to 70 years old. Neuropsychology: Shimamura and Squire (1986, 1988) examined metamemory in neuropsychological populations such as Korsakoff patients and chronic alcoholics, and Janowsky, Shimamura, and Squire (1989) examined metamemory in frontal-lobe patients. Special education: S. Krinsky (1988) examined metamemory in the deaf. Problem solving: Metcalfe (1986) examined metacognitions for solving problems, and Metcalfe and Wiebe (1987) discovered dynamic changes in metacognitions during problem solving. Psychopharmacology: Nelson, McSpadden *et al.* (1986) examined the effects of acute alcohol intoxication on metamemory. Naturalistic settings: Nelson *et al.* (1990) examined the way that metamemory is affected by extreme altitude at Mount Everest.[7]

V. Concluding Remarks

As mentioned earlier, our research in metamemory was initiated by the "paradoxical" findings that people can accurately predict their subsequent likelihood of recognizing nonrecallable items and that they can quickly and accurately decide—on the basis of no more than a cursory search through memory—that they will not retrieve particular sought-

[7]The pattern of results obtained at extreme altitude was in several ways the opposite of the pattern obtained from the alcohol-intoxicated people described in Nelson, McSpadden *et al.* (1986).

after items. Those findings led us to develop a methodology based on psychophysical methods that we used to empirically investigate people's "feeling of knowing." The results of our experiments convinced us that we were dealing with only a part of a complex metacognitive system and that to account adequately for feeling-of-knowing phenomena, a larger perspective was needed. This eventuated in the present theoretical framework that emphasizes the role of control and monitoring processes.

The embedding of the feeling of knowing in a richer framework helped to dissipate the paradoxical nature of the feeling of knowing. Moreover, in terms of our theoretical framework, an immediate goal of metamemory research is to explain the accuracy of metacognitive judgments in terms of remembered information, that is, to give a "No-Magic" explanation, like the one described above.

When we began 15 years ago, there were relatively few metamemory researchers and a paucity of solid empirical results about metamemory. We are pleased that today there are many capable, active investigators and a wealth of solid empirical findings. For our own work, we see the next big challenge to be the development of specific theories that will explicate the role of control and monitoring processes in human memory.

Acknowledgments

Without the continuous support of NIMH during the past 10 years, the research reported here would not have occurred. Preparation of this article was supported by NIMH grant MH32205. We thank Harry Bahrick for helpful comments on an earlier draft.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2). New York: Academic Press.

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, **77**, 215–222.

Bahrick, H. P., & Hall, L. K. (in press). Preventive and corrective maintenance of access to knowledge. *Applied Cognitive Psychology*.

Bower, G. H. (1967). A multicomponent theory of memory trace. In K. W. Spence (Ed.), *The Psychology of Learning*. New York: Academic Press.

Butterfield, E. C., & Belmont, J. M. (1971). Relations of storage and retrieval strategies as short-term memory processes. *Journal of Experimental Psychology*, **89**, 319–328.

Butterfield, E. C., Belmont, J. M., & Peltzman, D. J. (1971). Effects of recall requirement on acquisition strategy. *Journal of Experimental Psychology*, **90**, 347–348.

Butterfield, E. C., Nelson, T. O., & Peck, G. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology*, **24**, 654–663.

Carnap, R. (1934). *Logische syntax der sprache*. Vienna: Springer.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, **1**, 89–97.

Coombs, C. (1964). *A theory of data*. New York: Wiley.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, **87**, 215–251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, **34**, 906–911.

Groen, G., & Atkinson, R. C. (1966). Models for optimizing the learning process. *Psychological Bulletin*, **66**, 309–320.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, **56**, 208–216.

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

Hilbert, D. (1927). Uber das Unendliche. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, **36** 201–215.

Hirst, W. (1982). The amnesic syndrome: Descriptions and explanations. *Psychological Bulletin*, **91**, 435–460.

Hooker, W., & Jones, R. (1987). Increased susceptibility to memory intrusions and the Stroop interference effect during acute marijuana intoxication. *Psychopharmacology*, **91**, 20–24.

Jameson, K. A., Narens, L., Goldfarb, K., & Nelson, T. O. (1990). The influence of subthreshold priming on metamemory and recall. *Acta Psychologica*, **73**, 55–68.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Memory and metamemory: Comparisons between patients with frontal lobe lesions and amnesic patients. *Psychobiology*, **17**, 3–11.

Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics*, **10**, 8–14.

Kolers, P. A., & Palef, S. R. (1976). Knowing not. *Memory & Cognition*, **4**, 553–558.

Krinsky, R., & Nelson, T. O. (1985). The feeling of knowing for different types of retrieval failure. *Acta Psychologica*, **58**, 141–158.

Krinsky, S. (1988). *The feeling of knowing in deaf adolescents: A metamemorial study*. Doctoral dissertation, University of Washington, Seattle.

Lam, T. (1987). *An empirical investigation of extraneous factors and data collection procedures in the measurement of the feeling-of-knowing*. Doctoral dissertation, University of Washington, Seattle.

Le Ny, J. F., Denhiere, G., & Le Taillanter, D. (1972). Regulation of study-time and interstimulus similarity in self-paced learning conditions. *Acta Psychologica*, **36**, 280–289.

Leonesio, R. J. (1985). *Three measures of metamemory: Before you know, after you know, and after you don't know but feel you do*. Master's thesis, University of Washington, Seattle.

Leonesio, R. J., & Nelson, T. O. (1982). Postcriterion overlearning reduces the effectiveness of the method of adjusted learning. *Behavior Research Methods and Instrumentation*, **14**, 320–322.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same

underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

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

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, **74,** 100–109.

Loftus, G. R., & Wickens, T. D. (1970). Effect of incentive on storage and retrieval processes. *Journal of Experimental Psychology*, **85,** 141–147.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and cognition*, **10,** 663–679.

Marcel, A. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, **15,** 197–237.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory and cognition*, **18,** 196–204.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning , Memory, and Cognition*, **12,** 288–294.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, **15,** 238–246.

Modigliani, V., & Hedges, D. G. (1987). Distributed rehearsals and the primacy effect in single-trial free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13,** 426–436.

Nelson, T. O. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory*, **4,** 453–468.

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. (1986). BASIC programs for computation of Goodman-Kruskal gamma coefficient. *Bulletin of the Psychonomic Society*, **24,** 281–283.

Nelson, T. O. (1987). The Goodman-Kruskal gamma coefficient as an alternative to signal-detection theory's measures of absolute-judgment accuracy. In E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology*, Vol. 1, pp. 299–306 Amsterdam: Elsevier/North-Holland.

Nelson, T. O. (1988). Predictive accuracy of the feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. M. Gruneberg, P. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (Vol. 2). New York: Wiley.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). *Cognition and metacognition at extreme altitude on Mount Everest.* Manuscript under review.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, **113,** 282–300.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the 'labor-in-vain effect.' *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14,** 476–486.

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, and Cognition,* **12,** 279–287.

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, and Cognition,* **8,** 279–288.

Nelson, T. O., McSpadden, M., Fromme, K., & Marlatt, G. A. (1986). Effects of alcohol intoxication on metamemory and on retrieval from long-term memory. *Journal of Experimental Psychology: General,* **115,** 247–254.

Nelson, T. O., & Narens, L. (1980a). A new technique for investigating the feeling of knowing. *Acta Psychologica,* **46,**69–80.

Nelson, T. O., & Narens, L. (1980b). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior,* **19,** 338–368.

Nickerson, R. S. (1980). Motivated retrieval from archival memory. In *Nebraska Symposium on Motivation* (pp. 73–119).

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica,* **64,** 245–259.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* **84,** 231–259.

Pfefferbaum, A., Darley, C., Tinklenberg, J., Roth, W., & Kopell, B. (1977). Marijuana and memory intrusions. *Journal of Nervous and Mental Disease,* **165,** 381–386.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology,* **19,** 90–138.

Reder, L. M. (1988). Strategic control of retrieval strategies. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22). San Diego, CA: Academic Press.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology,* **39,** 475–543.

Schacter, D. L., & Worling, J. R. (1985). Attribute information and the feeling-of-knowing. *Canadian Journal of Psychology,* **39,** 467–475.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior,* **6,** 156–163.

Shimamura, A. P., Landwehr, R. F., & Nelson, T. O. (1981). FACTRETRIEVAL: A program for assessing someone's recall of general-information facts, feeling-of-knowing judgments for nonrecalled facts, and recognition of nonrecalled facts. *Behavior Research Methods and Instrumentation,* **13,** 691–692.

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **12,** 452–460.

Shimamura, A. P., & Squire, L. R. (1988). Long-term memory in amnesia: Cued recall, recognition memory, and confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **14,** 763–771.

Simon, H. A. (1979). *Models of thought.* New Haven, CT: Yale University Press.

Snow, J. (1987). *Design.* Cambridge, MA: Meta Software.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology,* **71,** 673–679.

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

Weingartner, H., Rudorfer, M. V., & Linnoila, M. (1985). Cognitive effects of lithium treatment in normal volunteers. *Psychopharmacology*, **86,** 472–474.

Wescourt, K. T., & Atkinson, R. C. (1973). Scanning for information in long- and short-term memory. *Journal of Experimental Psychology*, **98,** 95–101.

Wilkinson, T. S., & Nelson, T. O. (1984). FACTRETRIEVAL2: A Pascal program for assessing someone's recall of general-information facts, confidence about recall correctness, feeling-of-knowing judgments for nonrecalled facts, and recognition of nonrecalled facts. *Behavior Research Methods, Instruments and Computers*, **16,** 486–488.