COGNITIVE PSYCHOLOGY 19, 90–138 (1987)

Strategy Selection in Question Answering

LYNNE M. REDER

Carnegie-Mellon University

There are multiple strategies for answering questions. For example, a statement is sometimes verified using a plausibility process and sometimes using a direct retrieval process. It is claimed that there is a distinct strategy selection phase and a framework is proposed to account for strategy selection. Six experiments support the assumptions of the proposed framework: The first three experiments show that strategy selection is under the strategic control of the subjects. These experiments also indicate what contextual variables affect this selection. Experiments 4 and 5 suggest that strategy selection also involves evaluating the question itself, while Experiment 6 suggests variables that influence the evaluation of the question. This model is shown to be consistent with processing strategies in domains other than question answering, viz., dual-task monitoring in divided attention situations. © 1987 Academic Press, Inc.

In Norman's paper, *Memory, Knowledge, and the Answering of Questions* (1973), he points out that the process of question answering is far from simple and that the "traditional psychological studies of memory" do not tell us about the way that knowledge is used to answer questions. There has been considerable effort devoted to understanding memory and memory retrieval as it relates to recognition and recall tests. There are formal theories of how to find specific facts in memory (e.g., Anderson, 1972, 1976; Atkinson & Shiffrin, 1968; Bower, 1972; Kintsch, 1970; Raaijmakers & Shiffrin, 1981; Ratcliff & Murdock, 1976). There has also been work on other ways of answering questions, such as searching one's autobiographical memory (e.g., Reiser, Black, & Abelson, 1985; Whitten & Leonard, 1981; Williams & Hollan, 1981; Williams & Santos-Williams, 1980) and making plausibility judgments (e.g., Collins, 1978a, 1978b). Nickerson (1977a, 1977b, 1980a, 1980b) has argued that sometimes we retrieve information automatically from memory while other

This work was supported by Grant BNS-03711 from the National Science Foundation, and in part by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract N00014-84-K-0063, Contract Authority Identification Number, NR667-529. I thank J. Anderson, M. Bassok, D. Charney, A. Corbett, R. Jeffries, and J. McClelland for commenting on the manuscript. S. Fabri helped conduct Experiment 6. I am especially indebted to C. Dennler for invaluable assistance with all phases of the endeavor. Requests for reprints should be sent to Lynne M. Reder, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 15213.

0010-0285/87 $7.50

times retrieval is very effortful and purposeful. Despite all of this work, there has been little work on whether people intentionally select strategies and, if so, how people decide which strategy or process to use in order to answer a question.

Many question-answering models of memory include a preliminary stage where the subject performs an evaluation of the query to see if a quick decision can be made or if more work is required. For example, Norman (1973) pointed out that people do not search memory for the answer to the question, "What is Charles Dickens' telephone number?" Rather, some initial preprocessing allows us to decide that further search would be fruitless. A key assertion in this paper is that people's flexibility in their control of memory retrieval goes far beyond simply a decision whether to "fast exit." People have multiple strategies for retrieving information and choose to apply these strategies in variable orders. In a theory where there is variable strategy selection, it seems reasonable to propose that people's initial evaluation of their memory, with respect to the question, affects that strategy selection. There have been a number of endeavors concerned with self-assessment of knowledge (e.g., Gentner & Collins, 1981; Hart, 1965; J. L. Lachman & Lachman, 1980; Nelson, Gerler, & Narens 1984; Norman, 1973), although this work has not tended to address strategy selection.

This paper presents a general framework for the process of answering questions from memory. The paper focuses primarily on verification and recognition tasks, but not exclusively, and the framework is shown to generalize to other types of question-answering situations. There are a number of assumptions to the model proposed here. Before reviewing evidence for some of the claims and presenting new data in support of additional hypotheses, it would be useful to highlight the critical ideas:

1. There are multiple strategies for question answering.
2. One strategy is to try to find a fact which encodes the answer in memory. This is the direct retrieval strategy.
3. Another strategy is to compute a plausible answer given a set of facts stored in memory. This is the plausibility strategy.
4. Before answering a question, a person engages in an initial strategy-selection phase in order to decide which strategy or sequence of strategies to use.
5. The strategy-selection stage consists of an initial evaluation of knowledge relevant to the question followed by a decision of which strategy to follow. The initial evaluation is an automated process, while the decision is a controlled process.
6. In the initial evaluation, the person assesses how familiar the words

in the question are. The more familiar the words, the more the person is biased toward direct retrieval.

7. In the initial evaluation, the person also assesses how many intersections in memory there are among the words from the question. The more intersections, the more the person is biased toward plausibility.

8. The strategy decision process integrates information from the initial evaluation with factors extrinsic to the question in order to select a strategy. Extrinsic influences include instructions and probability that a particular strategy will be successful.

After reviewing the evidence that strategy selection (or bias) is involved whenever a question is answered, the paper goes on to suggest the mechanisms that people use for deciding quickly which strategy to apply. Understanding these mechanisms and the variables that influence them is the primary focus of this paper.

ARGUMENTS IN SUPPORT OF STRATEGY SELECTION

Searching Memory for a Specific Fact Is Not Always Preferred

The default assumption of memory theorists has tended to be that when verification of a proposition is required, a careful search for a specific proposition is the first strategy tried (at least after an initial evaluation).[1] The reason that direct retrieval is assumed to be tried first is that it is also commonly believed that direct matching is a more efficient process than inferential reasoning (e.g., Anderson, 1976; Anderson & Bower, 1973; Camp, J. L. Lachman, & Lachman, 1980; Collins & Loftus, 1975; Collins & Quillian, 1969; Haviland & Clark, 1974; Kintsch, 1974; J. L. Lachman & Lachman, 1980; R. Lachman, 1973; Lehnert, 1977; Norman, Rumelhart, & the LNR Research Group, 1975; Quillian, 1968; Schank & Abelson, 1977). J. L. Lachman and Lachman (1980) articulate this commonly held conception of the preference for one strategy over another:

> When a person needs a particular piece of information—e.g., to answer a question—she attempts to retrieve it directly. Metamemorial processes return the information that an answer is or is not in store. If an answer is found, metamemorial control processes are involved in assessing its adequacy. If no answer, or an inadequate answer, is retrieved, then the process of inference is set into motion. (pp. 289–290)

[1] For example, it is common for memory theorists to assume that Statement A was inferred during the reading of a text if verification of it is faster than verification of a different Statement B, (e.g., Garnham, 1982; Singer & Ferreira, 1983). Another interpretation, of course, is that Statement A is more plausible than Statement B and therefore faster to verify (judge plausible) at test, regardless of which statements had been inferred earlier.

There is now a growing body of literature suggesting that searching memory for an exact match is not always done even in tasks that require such careful inspections (e.g., Erikson & Mattson, 1981; Reder, 1982; Reder & Ross, 1983; Reder & Wible, 1984). Erikson and Mattson asked subjects questions like, "How many animals of each kind did Moses take on the Ark?" Subjects almost uniformly reply "two" even though they know that Noah took the animals on the ark. It seems in this case that people do not bother to carefully inspect their memories for exact matches to the memory probe. Their data cannot be explained by assuming that subjects accessed the correct memory trace, noted the discrepancy, but then obligingly gave the intended answer. If that were true, then subjects should not find it difficult to verbally note the discrepancy when specifically instructed to do so. In fact subjects have a great deal of difficulty with such a task: Reder and Dennler (1984) constructed a large number of these trick questions and told half the subjects to give the answer only when the question was presented in its correct form (i.e., answer when the question uses "Noah," but say "can't say" when the question uses "Moses") and told the other half of the subjects to give an answer based on the "gist" of the question (i.e., regardless of whether or not the question used "Moses"). Subjects were significantly faster and more accurate in the condition where they could ignore whether the question was properly formed or not. Subjects found it very difficult to say "can't say." It seems, then, that question answering often proceeds by loose inspection of the data base rather than by searching for one specific proposition.

The robustness of the "Moses illusion" suggests that subjects rarely prefer a strategy that involves a careful match to memory of one specific fact; however, other data suggest the opposite. For example, Singer and Ferreira (1983) found that subjects were faster to answer questions that involved an exact restatement of a sentence read in a story or that involved inferences likely to be drawn during reading than they were to answer sentence paraphrases or inferences not required for story comprehension. The evidence that subjects tend to verify statements by searching for an exact match comes from experiments involving short delays between study and test. In situations where the delays are longer, there is evidence that strategy preference changes from searching for exact matches to using a plausibility strategy (Reder, 1982; Reder & Wible, 1984).

Strategy Preference Is Not Stable for the Same Questions in the Same Task

In Reder (1982), subjects answered questions based on short stories they read. One group of subjects was required to decide whether a particular sentence had been studied, while the other group was to judge whether a particular statement was plausible given the story read. (See Table 1 for an example story.) Half of the plausible test probes had been presented in the story (a different, random set for each subject). Although it might seem reasonable that the verbatim or direct retrieval strategy would be used exclusively for the recognition task and the plausibility strategy for the plausibility task, the data indicated that subjects often tried first the strategy that corresponds to the other task. At short delays between reading of the story and test, subjects in both groups tended to prefer the direct retrieval (or verbatim match) strategy, while at longer delays, both groups tended to prefer the plausibility strategy.

Assessing strategy use. The evidence for use of both strategies in both tasks can been seen in several ways: The plausibility of the test probe produced latency differences and accuracy differences even in the recognition task; whether the test probe had been stated in the story or not caused latency and accuracy differences even when subjects were asked to judge whether the statement was plausible given the story (not whether it had been presented). This is not to say that the data look the same for subjects regardless of the official task: The plausibility effects are much larger for subjects asked to make plausibility judgments. And of course, whether the test probe had been presented in the story has a bigger effect on time and accuracy in the recognition task than in the plausibility task.

It is important to understand how strategy use is inferred, because it also is used to interpret new experiments described here. Test questions used in both the recognition task and the plausibility task were classified into types, half highly plausible, half moderately plausible (the plausibility task also had implausible questions). It makes sense that subjects take longer to decide that a moderately plausible statement is plausible than one classified as highly plausible. It is also reasonable that they less consistently judge moderately plausible statements to be plausible than highly plausible ones. These effects, however, should not obtain when subjects are asked to decide whether a statement was presented in the story unless subjects are sometimes making recognition judgments by judging plausibility. In fact, these plausibility effects do occur when subjects are asked to recognize whether or not a fact had been stated in the story: Moderately plausible statements were "recognized" more slowly than highly plausible; also, moderately plausible statements were "recognized" less often than highly plausible statements, regardless of whether or not they had actually been presented. This means that recognition accuracy is better for highly plausible statements than for moderately plausible statements when both types were presented in the story and worse for highly plausible statements when both plausibility types were not presented. In sum, latency and accuracy differences due to the plausibility of

TABLE 1
Example Story

We are out of touch with problems which were central in the past.
But this is not true everywhere.
The setting is Burma.
The tiger was a man-eater.
It suffered from an old gunshot wound.
The villagers did not dare work in the fields.
In a disorderly meeting they made a decision.
They asked a hunter to help them.
He came the following week.
The tiger killed a man.
It had attacked him in a small ravine.
It carried the victim away.
It concealed the kill under some vines.
The hunter followed the tiger's trail.
The traces were distinct.
He found the tiger asleep.
The shade was cool there.
The hunter considered giving the tiger a sporting chance.
Then he shot it.
The tiger died quietly.
The hunter did not feel right.
The villagers understood his feelings.
But their concern was practical.
The hunter skinned the tiger.
He left with the skin.

Statements to judge[a]

Highly plausible
The tiger was dangerous.
The villagers were afraid of the tiger.
The villagers were happy that the tiger was dead.

Moderately plausible
The hunter was expected to solve their problem.
The hunter thought of the tiger's situation.
The hunter used his best judgment.

Implausible
The villagers had encountered many man-eating tigers.
The hunter was well paid for killing the tiger.
The villagers make their living primarily by hunting.
There are no guns in Burma.
The villagers are Hindu and do not believe in killing.
The village chieftain wore the tiger skin over his hips.

Contradictory
The tiger was awake when the hunter found him.
The villagers live in Nepal.
The hunter left the tiger's pelt with the villagers.
The tiger died fighting.
The villagers scorned the hunter's feelings.
The tiger concealed his victim in a cave.

[a] Although 6 implausible and 6 contradictory statements are listed here, only 6 of the possible 12 were selected for any story. This ensured an equal number of true and false test statements.

the test item are evidence of the use of the plausibility judgment strategy. and it seems clear that this strategy is sometimes used to make recognition judgments.

Latency and accuracy differences due to whether a probe has been presented in the story provide converging evidence for claims about strategy use. For subjects asked to make plausibility judgments. plausible statements that had not been presented in the story were more often judged "implausible." This suggests that subjects were sometimes using the direct-retrieval strategy when asked to make plausibility judgments. Consistent with this result on accuracy, the differences in RT between moderately and highly plausible statements in the plausibility task was larger for probes that had not appeared in the story: These statements could not be verified by the direct retrieval strategy.

Evidence for strategy shifts. As the delay between study and test increased, there was a shift away from use of the direct retrieval strategy toward greater use of the plausibility strategy. This evidence comes from the change in the size of plausibility effects with delay. Differences in RT between moderately and highly plausible statements increased with delay for presented statements.[2] This suggests that people change strategy preference from direct retrieval at short delays to the plausibility strategy at longer delays.

The increase in error rates in the recognition task for not-stated items, especially for highly plausible statements, also indicates a shift in preference for the plausibility strategy with longer delays: At longer delays, subjects tend to prefer the plausibility strategy even in the recognition task. Highly plausible statements are more likely to be judged plausible, causing an error for those statements that had not been presented in the story (70% errors for highly plausible not stated; 50% errors for moderately plausible not stated).

Converging Results That Support a Strategy-Selection Stage

The interpretation that strategy choice varies with the delay between reading the story and test suggests that people have a mechanism that allows them to select a strategy for question answering prior to executing that strategy. Below I review additional data that support a preliminary strategy-selection phase. In addition, new experiments are reported that further strengthen the case for an initial selection phase.

By assuming that people are able to select the strategy that they use in tasks such as question answering, a number of results are more easily interpreted. For example, an unusual finding from Reder (1982) was that

[2] For statements that had not been presented. plausibility had to be used for correct responding. Therefore, strategy shifts are less noticeable.

subjects asked to make a plausibility judgment were very slow to judge not-stated items as plausible, but only when the delay between reading and test was short. As the delay increased, subjects actually became faster than they were at shorter delays to judge not-stated items as plausible.

The explanation given in Reder (1982) was that at short delays, the wrong strategy is tried first. The flowchart model displayed in Fig. 1 represents the probabilistic model offered in that paper. It represents the branching alternatives associated with judging an assertion, regardless of whether the person was asked to make a plausibility judgment or a recognition judgment. Each branch (reflecting a choice path) has a probability associated with it and is affected by variables such as the delay between

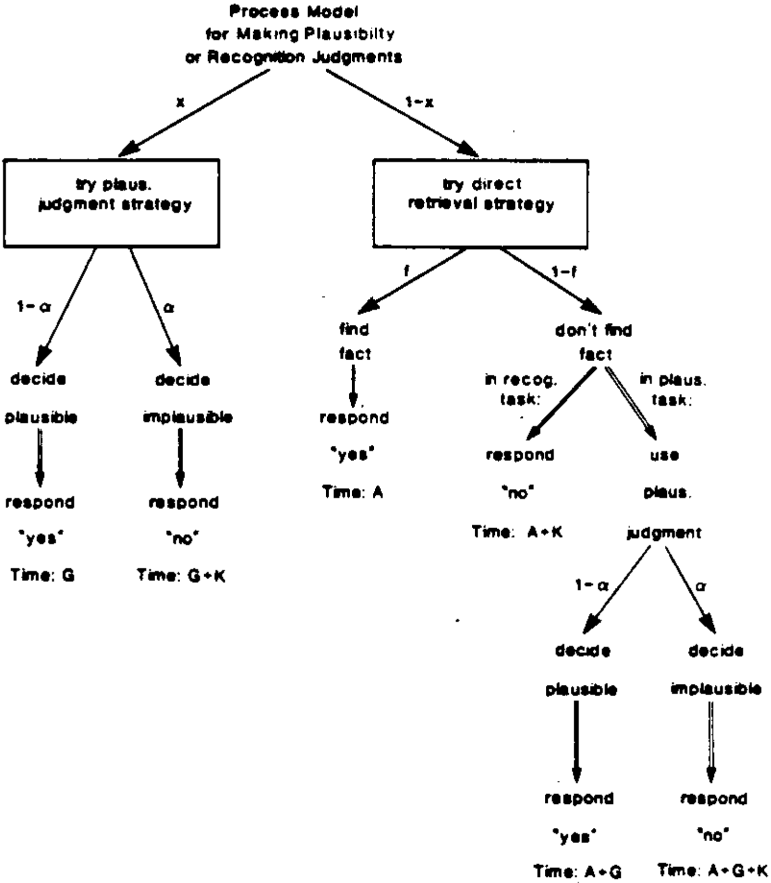

FIG. 1. Flowchart model of strategy selection from Reder (1982).

reading the story and test, the official task requirements or the plausibility of the test probe.

The speedup for not-presented plausible statements was explained by assuming that at short delays subjects tried the direct retrieval strategy first (right branch of tree), and when it failed, they went on to compute the statement's plausibility; at longer delays, they tried plausibility without first executing the useless strategy (left branch), thereby saving time.

Put another way, the claim is that people sometimes adopt as a first strategy one that is inappropriate for certain conditions. This can also explain a speedup found in the data of Reder and Wible (1984). That experiment required subjects to make either recognition judgments or consistency judgments about statements after having studied groups of thematically related facts. Judging consistency meant deciding whether a probe was thematically similar to a studied statement. At short delays, subjects were very slow to verify not-stated, consistent items in a consistency judgment task. At longer delays, they became faster and more accurate to verify this type of item, while they became much slower to reject those items in a recognition task. Again, the explanation was that subjects preferred the direct retrieval strategy at short delays–an inappropriate strategy for not-stated, consistent items in the consistency judgment task. At longer delays, the direct retrieval strategy was often avoided, saving time for not-stated items in the consistency judgment task, but causing errors or slow responses for not-stated consistent items in the recognition task.

Other Factors That Influence Strategy Selection

The tendency to prefer the direct retrieval strategy or the consistency strategy was not just affected by delay between study and test. It was also influenced by whether the official task was to make recognition judgments or consistency judgments. Even in situations where there are not explicit instructions, strategy selection can be affected by impressions or expectations. For instance, Gould and Stephenson (1967) found that subjects' willingness to engage in reconstructive recall was affected by their perception of the emphasis on verbatim recall.

Reder and Ross (1983) have shown that even within a recognition task, strategy preference can depend on the relation between the true and false assertions to be discriminated, i.e., subjects adjust their strategy selection to reflect the difficulty of the discrimination. In Reder and Ross, subjects studied related sets of facts about fictitious individuals (e.g., Marty going to the circus), and had to discriminate studied sentences from nonstudied foils. When these foils were thematically related to the studied facts (e.g., also about Marty at the circus), subjects tended to adopt the direct retrieval strategy. When foils were unrelated to the circus theme, subjects tended to base their "recognition judgments" on a plausibility strategy.

In a similar vein, Lorch (1981) showed that in a category membership task, when unrelated items were used as foils, subjects tended to adopt a strategy that seemed consistent with the semantic overlap model of Smith et al. When foils consisted of highly related and somewhat related terms, but no unrelated terms, subjects tended to adopt a strategy of careful evaluation of subject–predicate relations.

Serial or Parallel Processes?

Past experiments described above, as well as ones reported here, argue strongly for the existence of a strategy selection process. It seems natural to assume that this selection process occurs prior to strategy execution, although one could argue that strategy selection and strategy execution operate in parallel. Indeed, within this model there are two levels at which processes could operate serially or in parallel: Strategy selection could operate before execution of a strategy or at the same time, and the execution of the available processes could be serial or the competing processes could operate in parallel. Townsend has shown (e.g., 1971, 1974, 1976; see also Ross & Anderson, 1981; Snodgrass & Townsend, 1980, for further discussion), that for any particular specification of a serial model, a specific parallel model can be constructed that produces the same results.

Let me briefly describe four different ways strategy selection and execution could operate, depending on whether one assumed that strategy selection and strategy execution co-occur or are deployed serially, and whether the biased strategies execute serially or in parallel.

- One model (serial/serial) assumes that strategy selection must be complete before strategy execution starts. The selected strategy executes by itself and tries to find the answer. (The less favored strategy might execute afterwards.)
- Another model (serial/parallel) would also assume that there is a prior strategy selection phase but instead of the selected strategy executing alone, it just receives greater processing capacity than the less preferred strategy that executes with it, in parallel.
- It is also conceivable that the strategy-selection process operates while a strategy is executing (parallel/serial). If the selection process completes before the executing strategy, and the selection process prefers the other procedure, then the executing strategy is terminated and the other strategy starts up.

- An example of the remaining case (parallel/parallel) is one where the selection process operates at the same time as both strategies which are racing with equal processing capacity. The outcome of the strategy-selection process is to "gate" or inhibit output from the less ' vored strategy.

In Reder (1982), I presented a model in which subjects sequentially tried one strategy and then another; however, it was noted there, too, that an equivalent model would involve a (parallel) race between the two strategies where subjects biased the amount of cognitive capacity given to each strategy. For purposes of the present paper I do not care whether execution of strategies occurs in a non-fixed-order serial manner or in parallel, with differential allocation of resources. On the other hand, in this paper, evidence that supports the existence of a strategy selection process is interpreted as a process that occurs prior to the execution of the strategies. This seems like a simpler and more natural way to understand how we answer questions.

NEW EVIDENCE FOR CONTROLLED STRATEGY SELECTION: THE ROLE OF SITUATIONAL VARIABLES

The data reviewed thus far seem easily explained if one assumes that people have the ability to decide how they want to go about answering questions. This assumption leaves open a number of questions. For example, does a person decide each time he or she attempts a question which strategy will be preferred? Or do people select a preferred strategy to use throughout a task based on knowledge of instructions and similarity of foils to targets? How sensitive are people to the success rate of a particular strategy? Can we quickly adjust strategy preference based on subtle features such as success with a strategy?

This section describes experiments that are concerned with the extent to which people can fine-tune their control over what strategy they use for question answering. It is also concerned with uncovering what factors extrinsic to the question affect strategy-selection. These experiments also provide further support that people can and do select a strategy prior to executing one for question answering.

EXPERIMENT 1

Can We Adjust Our Strategy Preference to Mirror the Ratio of Presented to Nonpresented Statements in a Story?

To what extent can people discern the effectiveness of a particular strategy and adjust the tendency to select a strategy on the basis of its perceived effectiveness? To address this question, subjects read short, mildly interesting stories and then were asked questions about them. After each story, subjects were asked to judge whether the test probe was plausible, given the story. Unlike previous experiments of this type, the percentage of explicitly presented inferences varied from the usual 50%. For half of the subjects, 80% of the plausible statements were presented in the stories, and for the other half of the subjects only 20% were presented. Of interest were whether the propensity to use a strategy was affected by the different ratios of presented to not-presented sentences, and how quickly subjects adjusted their strategies to adapt to these ratios.

Method

Materials. Ten stories written by five different authors were used. The questions about the stories and the stories themselves had been used previously (Reder, 1976, 1979, 1982). The questions were of three types: highly plausible, moderately plausible, and implausible. The plausibility of plausible statements had been determined by previous subject ratings. Half of the implausible statements were contradictions. (Contradictions had not been used in the previous studies.) Each contradictory statement was an exact restatement of a statement from the story except that one word was replaced by its opposite. Contradictory statements were considered "presented" implausibles. The implausible and contradictory statements did not vary systematically on implausibility. They were constructed by the experimenter with the constraint that noncontradictions not refer to any specific statement in the story.[3] Table 1 gives an example story with implausible and contradictory statements.

Design and procedure. There were four factors in this experiment: whether the ratio of presented statements to nonpresented statements favored the direct retrieval strategy or the plausibility strategy, whether the statement itself was highly or moderately plausible, whether the statement had actually been presented in the story, and whether the bias was still present. This last factor was manipulated by the following design of the materials: The first 6 of the 10 stories had either the 80/20 split or the 20/80 split of presented to not-presented plausible statements; however, after the first 6 stories, the remaining 4 stories always returned to the more conventional 50/50 split.

The experiment was conducted on an IBM personal computer. Subjects read 10 stories which were titled and presented in random order. For each story, subjects controlled the rate at which statements appeared on the screen: each time the space bar was pressed, a new statement appeared, so long as the previous statement had been on the screen for a minimum of 0.5 sec. Subjects were tested after reading each story. Subjects were asked to decide whether or not each statement was plausible, i.e., consistent with the information in the story just read. Subjects were told to indicate that a statement was plausible or implausible by pressing the "K" or the "D" key, respectively. They were further instructed to keep their index fingers on these keys at all times while judging the statements, since response times would be recorded, and to respond as quickly as possible, without sacrificing accuracy.

Subjects. College-age subjects who read ads on the Carnegie-Mellon and University of

[3] Just as a presented plausible can be verified by either strategy, a contradictory statement can be rejected by either strategy, viz., finding its exact contradiction in memory and noting the discrepancy or by inferring that it is implausible. Note that failing to find a fact using the direct retrieval strategy is not a valid basis for deciding that a fact is implausible.

Pittsburgh campuses were recruited. Thirty subjects were randomly assigned to the Inference bias condition (i.e., only 20% of the plausibles were stated in the stories for the first six stories) and 29 to the Direct retrieval bias condition (i.e., 80% of the plausible probes had been presented in the stories).

Results

Below in Table 2 are listed the mean response times to make plausibility judgments as a function of four factors: the plausibility of the statement, whether the statement had been presented or not, the direction of bias (80% presented—biasing direct retrieval versus 80% not presented —biasing plausibility), anđ whether the ratio of presented to not-presented had reverted to 50:50. Medians of correct response times in each condition for each subject were computed. The times presented here represent the mean of these medians. If a subject had no correct response times in a condition, the cell was estimated by taking the grand mean and subtracting from it both the effect size for that condition and the effect size for that subject. Less than 0.003 of the cells needed to be estimated in that fashion. Figure 2 presents the difference between moderately and highly plausible response times as a function of biasing condition, collapsed over the stated/not-stated factor. It is plotted for Stories 1–6,

TABLE 2
Mean Response Times (and Accuracy) to Make Verification Judgments in Experiment 1

	Biased for direct retrieval (80% stated)		Biased for plausibility (80% not stated)	
	Stated	Not stated	Stated	Not stated
		Biased: Stories 1–6		
Highly	2.12 (0.92)	3.13 (0.88)	2.56 (0.90)	2.60 (0.90)
Moderately	2.27 (0.90)	3.13 (0.77)	2.84 (0.92)	3.21 (0.74)
Implaus/contra	2.57 (0.82)	2.80 (0.84)	3.11 (0.86)	2.96 (0.89)
		Returned to 50% stated: Stories 7–10		
Highly	2.15 (0.95)	2.31 (0.91)	2.30 (0.98)	2.57 (0.95)
Moderately	2.31 (0.94)	3.02 (0.73)	2.33 (0.96)	2.97 (0.82)
Implaus/contra	2.31 (0.86)	2.49 (0.92)	2.61 (0.89)	2.90 (0.90)

Note. The implausible statements were never presented in the story. The contradictory statements contradicted an explicitly presented statement by substituting an opposite for one word in the statement.

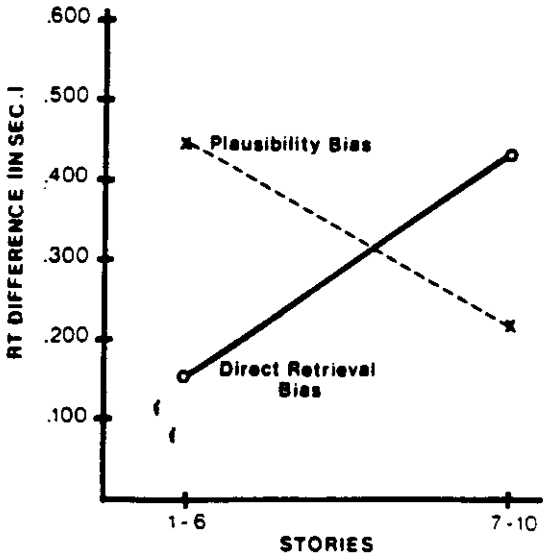

FIG. 2. Difference in RT between highly and moderately plausible statements (collapsed over stated/not stated) as a function of strategy bias, when bias imposed (Stories 1–6) and not imposed (Stories 7–10) in Experiment 1.

where bias existed, and Stories 7–10, where the ratio reverted to 50:50. Figure 3 presents the difference between stated and not stated, collapsed over plausibility.

There are a number of interesting patterns worth noting. Differences were plotted rather than the raw data, because this way it is easier to see how dramatic the effects are. Figure 2 plots the extent to which there was an effect due to plausibility of the question (moderately plausible RT minus highly plausible). The effect due to plausibility of the question was much greater for subjects biased to use the plausibility strategy. This was true both for questions that had been presented in the story and for those that had not. The results shown in Fig. 3 are in stark contrast. Figure 3 plots the extent to which there was an effect due to whether the question had been presented in the story (not-stated minus stated). In this case, there was a large effect for subjects biased to use the direct retrieval strategy—just the opposite of Fig. 2, which showed the plausibility effect. These different trends for the two groups are exactly what one would expect if the ratio of questions previously presented to not previously presented in the story actually did bias subjects' preference for a particular strategy.

An ANOVA was performed on the data from the first six stories to determine whether the contrasts mentioned above were significant. For brevity, the reporting of standard results, e.g., main effects due to

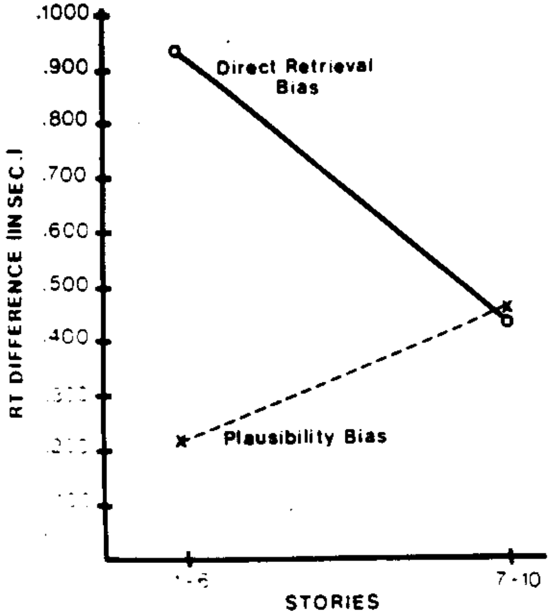

FIG. 3. Difference in RT between stated and not-stated statements (collapsed over plausibility) as a function of strategy bias, when bias imposed (Stories 1–6) and not imposed (Stories 7–10) in Experiment 1.

whether the probe was stated or not, or the plausibility of the probe, are omitted. All expected replications were obtained. There was a significant interaction on RT between bias (whether 80% of the probes were stated or only 20% were stated) and whether or not the probe was stated in the story, $F(1,57) = 16.1, p < .01$. This interaction represents the finding that the difference between stated and not-stated items was greater in the condition where the ratio of items biased subjects to adopt the direct retrieval strategy.[4]

The interaction of plausibility with bias in strategy selection also significantly affected RT, $F(1,57) = 6.2, p < .05$ for the first six stories, such that the differences between highly and moderately plausible statements was larger for subjects who were biased to use the plausibility strategy. (There was no interaction for the last four stories.)

There were also some interesting results with respect to how subjects readjusted to a 50/50 split and how easily they altered their strategies. For example, there was a significant interaction of Stated × Bias × "Half" (first 6 stories vs last 4), $F(1,57) = 11.8, p < .01$. That statistic represents

[4] Note that the difference in response times is not due to different responses to not-stated items: Both groups are expected to respond affirmatively to not-stated plausible probes.

the fact that when the bias manipulation stopped, the stated/not-stated difference decreased for subjects originally biased to use direct retrieval and increased for subjects originally biased to use plausibility. The effect became equivalent for the two groups. Another triple interaction which illustrates the same idea was the interaction of Probe Plausibility × Bias × "Half," $F(1,57) = 8.1, p < .01$. Originally the plausibility effects were much bigger for subjects biased to use the plausibility strategy, but they got smaller when the bias disappeared. Conversely, the small plausibility effect for subjects biased to use direct retrieval increased when the bias disappeared. Although the plausibility effect appears to have reversed itself when the bias was removed, the Plausibility × Bias interaction was not significant for the last four stories, $p > .10$.

Discussion

In this experiment, the official task, judging plausibility, was identical for both groups of subjects, so all differences in performance were due to the effectiveness of a particular strategy. These results strongly suggest that people are sensitive to the parameters of the situation in which they find themselves and can rapidly alter the strategy they employ. In past experiments, subjects were shown to adjust their strategy as the delay increased, perhaps because they knew that the retrieval strategy would be less effective at a delay. Past studies also showed that preference for a strategy depends on what subjects are actually asked to do, viz., make recognition judgments or make plausibility judgments. The next study examines whether that result is due to official demands, per se, or only to subjects' sensitivity to the probability of success with a strategy in a given task.

EXPERIMENT 2

What Strategy Is Preferred When Both Strategies Always Work?

It is conceivable that strategy selection would not be affected by official task instructions if either strategy worked equally well for the required task. In past studies (e.g., Reder, 1982), where subjects were asked to make recognition judgments, they were more inclined to use the direct retrieval strategy. That greater tendency to select the direct retrieval strategy may have occurred because they did not want to make all the errors that would result from adopting the plausibility strategy, and not because of the official task demands, per se. For this reason, I conducted a study where either strategy could apply equally well, and looked to see the effects of official task demands and delay on strategy selection.

In this study, all plausible statements were presented in the story and no implausibles were presented. Therefore, subjects could use either strategy and always be correct (assuming that they did not forget the statement or make an erroneous plausibility judgment). Nonetheless, I expected subjects to show less use of the plausibility strategy when instructed to make recognition judgments. For example, the latency differences between highly and moderately plausible statements should be smaller when asked to make recognition judgments.

Method

Design and materials. The general design and materials were similar to Experiment 1 with several important differences. Half of the subjects were randomly assigned to make recognition judgments and the other half to make plausibility judgments. Instead of varying the proportion of plausible statements included in the story, all plausible statements to be judged about a story had been included as part of the story. Half of the subjects assigned to each task were tested after each story and the other half of each task were tested after reading all 10 stories.

The stories and test statements were the same as in Experiment 1 except that all plausible test items were presented in the story. None of the implausible statements were contradictions. In this way, subjects asked to make recognition judgments could use the plausibility strategy without making errors and subjects asked to make plausibility judgments could use the direct retrieval strategy without making errors.

Procedure. The experiment was quite similar to the procedure of Experiment 1, with only a few relevant changes. Depending on condition, subjects were either told that they would be questioned after reading each story or after reading all 10 stories. Each story's test statements were preceded by the story's title in the Delay condition. Subjects assigned to the Recognition condition were asked to decide whether or not each statement was exactly the same as one that they had read in the story, and not to respond affirmatively because a statement seemed true, if it had not actually been read.

Subjects. Sixty-two subjects were recruited from the department's summertime subject pool and randomly assigned to conditions. In the Immediate condition, there were 17 subjects asked to make recognition judgments and 14 asked to make plausibility judgments. At the longer-delay (approximately 20 min. after reading all 10 stories), there were 16 assigned to the recognition task and 15 to the plausibility task. Subjects received $4.50 for participating both in this 30- to 40-min experiment and in one other that followed.

Results and Discussion

Table 3 presents the data from Experiment 2, organized by delay, task, and plausibility of the statements. These data are the means of subjects' correct median response times and the proportion of correct trials per condition. Analyses of variance were performed on the median correct response times and accuracy data using the same factors mentioned above. Analyses were done using different contrasts of plausibility: plausible vs implausible, and highly plausible vs moderately plausible.

Note that at the short delay interval, there was a 0.115-s advantage for the highly plausible statements over the moderately plausible statements in the plausibility task. This difference grew to a 0.266-s advantage at the longer delay, an increase of 0.151 s. When subjects were asked to make recognition judgments, initially they were actually slightly slower (less than 0.025 s) for highly plausible statements than for moderately plausible statements, but here, too, the tendency to adopt the plausibility strategy

TABLE 3
Mean Response Times and Accuracy for Judgments in Experiment 2

	Recognition		Plausibility	
Official task	% Correct	RT	% Correct	RT
Immediate				
Med. stated	90.56	2.103	93.00	1.964
High stated	93.90	2.125	94.47	1.849
Implausible	96.06	2.178	89.86	2.300
Delayed				
Med. stated	88.54	1.979	91.09	2.320
High stated	90.00	1.889	94.89	2.054
Implausible	91.24	1.983	88.87	2.519

increased: At the delayed test, the advantage of the highly plausible grew to 0.090 s, an increase of nearly 0.115 s. The contrast using only plausible statements showed a marginally significant growth in the plausibility effect for both task types, $F(1,58) = 2.8$, $p < .10$.

The ANOVA that contrasted highly with moderately plausible statements produced a significant interaction of plausibility and task on response times, $F(1,58) = 4.0$, $p < .05$, such that the difference between highly and moderately plausible statements was bigger for subjects actually asked to make plausibility judgments. The comparison of plausible statements with implausible statements also interacted significantly with task instructions for both accuracy and response times, $F(1,58) = 16.9$ and 15.8, respectively, $p < .01$: Apparently, it is easier to say that an implausible statement was not presented (both in terms of RTs and accuracy) than to judge it as implausible.

These data make clear that strategy-selection preference is affected by official task requirements even when the strategy selection has no impact on performance. A different way of putting it is that subjects do not just follow instructions because they do not want to make errors. Still, the delay between study and test also seemed to influence strategy selection. The present experiment cannot tell us, however, whether the shift in strategy preference with delay is due to a conscious decision, i.e., being aware of the delay, or whether it is due to an impression that the information is now less available.[5]

EXPERIMENT 3

Can People Switch Strategies from Question to Question Based on Advice Preceding Each One?

The previous studies examined the differences due to instructions,

[5] This issue is discussed at length later in the paper.

e.g., asking subjects to make recognition or plausibility judgments. These instructions were given only once, prior to reading the stories. Subjects could develop a "frame of mind" and could keep the same bias for the duration of the experiment. Experiment 1 showed that subjects can and do shift their bias during an experiment, but the shift shown in that experiment was not from question to question. It is unclear whether subjects can consciously alter strategy selection from question to question, at a moment's notice. The ability to rapidly switch strategies might be fairly unconscious and not something a person can self-instruct in a matter of seconds.

In this experiment all subjects were asked to make plausibility judgments. However, before each question appeared on the computer screen, the subject was given "advice" by the computer as to which strategy was likely to be easier to use to answer the next question. For example, if the subject was to judge the plausibility of a statement that had been (recently) presented, the computer might advise that the subject "search memory" to try to find the relevant statement. If the statement had not been presented, the advice might be to try to "infer" the answer rather than searching for it directly. To make the advice useful, and in order to see whether the advice was having any effect, the advice was appropriate the majority of the time, but not always. On 80% of the trials, the advice was appropriate, and on 20% of the trials, the advice was inappropriate.

Method

Design and materials. The design consisted of the factors of probe plausibility (highly vs moderately vs not plausible), probe presentation in story (stated vs not stated), and advised strategy (inference vs direct retrieval of statement).Whether the advised strategy was appropriate or inappropriate depended on the recommendation and on whether the statement had been presented in the story.

The stories were the same 10 used in Experiments 1 and 2. The questions were also the same as those in Experiment 1 (three of the six implausibles per story contradicted a presented statement). Contradictory statements were considered presented, not-plausible statements.

Before each trial subjects received advice, half of the time to search for a specific fact, and half to try to judge whether the statement was plausible. Half of the plausible statements were presented in the story for all subjects and no implausible or contradictories were presented. It was still possible to design the program to have the advice be correct exactly 80% of the time. As in all studies, assignment of questions to condition was randomly determined, as was order of presentation of stories, with one exception: For the first story pair, only correct advice was given so that subjects would more rapidly learn to attend to the advice.

Procedure. The procedure was very similar to Experiments 1 and 2. Questions were asked about stories after every two stories. The first of a pair was questioned, then the second. The relevant part of the instructions said,

> Before seeing each statement, the computer screen will advise you to use a particular strategy to judge the plausibility of the statement. This advice is based on whether the statement or its contradiction was actually presented in the story.

> You will be told either "Try to retrieve a specific fact to use in judgment." or "Try to infer the answer." Most of the time the advice is going to be helpful. That is, when you are advised to retrieve a fact, either that fact or its contradiction was stated in the story. Similarly, when advised to infer, most of the time neither the fact nor its contradiction was stated. However, occasionally the advice will be wrong, such that the opposite advice would have been correct. Should the advice be wrong, still try to answer the question correctly.

The instructions continued with concrete examples, and advice about maintaining speed and accuracy and keeping index fingers on the response keys during the testing phase. The advice that preceded each question was displayed for a minimum of 0.5 s. Then hitting the space bar allowed the question to be presented.

Subjects. Eighteen undergraduates enrolled in psychology classes at C-MU participated for one credit toward a course requirement. One subject's data were lost due to technical difficulties with the computer.

Results and Discussion

Means of median correct response times and proportion of correct responses are displayed in Table 4 as a function of advice suggested, plausibility of the probe, and whether the probe had been stated in the story. The suitability or appropriateness of the advice is indicated by a (+) or (−). For example, for the not-stated items, the inference advice is appropriate and the direct retrieval advice inappropriate. Thus, the line above the not-stated items says "Inf.(+) dir. ret.(−)." The data from the first story pair are not included, because no incorrect advice was given while attempting to get subjects to attend to the advice.

An ANOVA was performed using the factors of probe plausibility, probe presentation, and strategy advised. The first thing to note is that subjects were significantly faster if the advice was correct, $F(1,16) = 5.59$, $p < .05$ (interaction of advice and probe presentation). That is, subjects were faster with the direct retrieval advice when the probes were

TABLE 4
Mean Response Time and Accuracy for Verification Judgments as a Function of Advice Type, Plausibility, and Whether the Probe Had Been Presented in the Story

	Inference advice		Direct retrieval advice	
	% Correct	RT	% Correct	RT
Inf.(+) dir. ret.(−)				
Med, not stated	74.54	2.706	82.35	3.281
High, not stated	92.81	2.268	91.18	2.214
Implausible	80.46	2.530	84.90	2.738
Dir. ret.(+) inf.(−)				
Med, stated	92.16	2.444	93.46	2.049
High, stated	98.04	2.101	99.35	2.001
Contradictory	84.12	2.683	76.41	2.501

not stated in the story and were faster with the inference advice when the probes were not stated in the story. The RTs for not-stated probes when direct retrieval was advised were significantly slower than all others, $t(32) = 3.82, p < .01$, because this was the only condition where the advised strategy would not work. (The advice to infer when the probe was stated will work; however, it is nonoptimal, since at such a short delay, direct retrieval is a faster strategy.) For highly plausible statements, the suitability of the advice had little effect. This suggests that many highly plausible statements were inferred by the subject when they had not been explicitly stated.

A second pattern worth mentioning is the difference in RT between the moderately and highly plausible statements as a function of strategy advised and whether or not the advice would work. The interaction of Stated × Plausibility (alternatively called Correctness of advice × Type of advice × Plausibility) was significant, $F(2,32) = 6.54, p < .01$. For stated probes, when searching for specific facts was advised, there was less than a 0.050-s difference between the moderately and highly plausible statements. For these same probes, the difference was over 0.300 s if inference was advised. For not-stated probes, there was a large difference due to plausibility regardless of strategy advised, since the inference strategy had to be executed ultimately. When direct retrieval was advised, the moderately plausible were very slow, since two strategies had to be tried; however, the highly plausible statements did not show this pattern. This again suggests that the highly plausible inferences were found in memory during the direct retrieval search (i.e., were inferred during reading), so that the second strategy did not have to be evoked. Suitability of the advice did not have a reliable effect on accuracy, but accuracy was affected by whether the probe had been stated in the story, $F(1,16) = 7.20, p < .05$.[6]

SUMMARY AND IMPLICATIONS

Experiments 1, 2, and 3 showed the extent to which strategy selection can be influenced by factors extrinsic to the test question. Since the ratio of presented to not-presented statements affects the probability of success of the direct retrieval strategy, subjects in Experiment 1 adjusted their use of strategies accordingly. Experiment 1 also showed that people quickly adapt to a change in this ratio. To balance Experiment 1, Experi-

[6] The only probe that had lower accuracy when "stated" in the story was the contradictory statement when direct retrieval was advised. Perhaps this results from subjects not bothering to note the one opposite word in the almost-verbatim match of the memory trace to the contradictory probe. That is, the direct retrieval strategy is a "literal" match strategy and sometimes subjects are sloppy and do not check every word.

ment 2 showed that subjects' strategy selection is affected by the official task instructions even when either strategy (direct retrieval or plausibility judgments) will produce the correct response. There were larger plausibility effects in conditions where subjects were actually asked to make plausibility judgments. The plausibility effects were larger at longer delays for subjects in both task conditions, confirming that other variables also influence choice.

The results from Experiment 3 give support to the idea that people can rapidly alter the strategy they use to answer a question. Subjects could modify which strategy they selected from question to question on the basis of an external cue such as "try to find the fact in memory" or "try to infer whether this statement is plausible." Plausibility effects were small if the direct retrieval advice was given and would work. The difference in RT between questions that had been stated and those that had not been stated was small if the advice was to use plausibility.

Experiment 3 clearly indicated that performance suffers when the wrong strategy is selected. Considering the results of this experiment, those of Experiments 1 and 2, and all the evidence reviewed earlier, it seems clear that any question-answering model requires a preliminary strategy selection phase. Given the existence of a strategy selection stage, it is interesting to ask whether factors besides situational or contextual variables play a role in this strategy selection. The distinction here is between variables extrinsic to the test question and variables intrinsic to the question. The next section addresses this issue.

THE ROLE OF STIMULUS EVALUATION IN STRATEGY SELECTION

There is reason to believe that, in addition to situational or extrinsic variables, variables intrinsic to the test question also play a role in strategy selection. Reder (1982), Reder and Wible (1984), and Experiment 2 all show that subjects shifted their strategy preference with delay. The explanation for the strategy shift was that at longer delays information is less accessible, making direct retrieval less desirable. It is possible that the decision to shift strategies was based on the subject's knowledge of the delay between study and test; however, I believe that subjects shifted strategy by evaluating the familiarity of the test questions. This is because in most (nonexperimental) situations, people do not know in advance when the queried information was learned and consequently need to have some mechanism for assessing familiarity and choosing a strategy.

The position that sentential or intrinsic variables should affect strategy selection is consistent with other views of question answering. A number of models have suggested that subjects make memory judgments by a two-stage process of (1) an initial memory evaluation, followed by (2) an optional, second process that more carefully inspects memory. For example, Atkinson and Juola (1974), in their analysis of word recognition, propose that subjects initially assess the "familiarity" of an item. If the item seems highly familiar (e.g., "I'm sure I've seen this recently"), they recognize the item; if it is of low familiarity they reject it. For intermediate levels of familiarity, subjects have to engage in a search of memory.

Smith, Shoben, and Rips (1974) proposed a similar idea for category judgments (e.g., "a chicken is a bird," "a canary is a bird"). They proposed that subjects evaluated the raw similarity (or relatedness) between subject and predicate. Again, if similarity was high, the statement was accepted; if low, it was rejected. For intermediate values, a careful inspection of the defining features was required, ignoring overlaps on "characteristic features."

Glucksberg and McCloskey (1981) also have evidence consistent with a preliminary stage that precedes careful inspection. In this case, the data suggested that subjects make a quick exit if the preliminary inspection fails to find a connection among the concepts. Subjects were much faster to respond, "It is unknown whether John possessed a gun" if they had not studied anything about John and guns than if they had studied that exact proposition, "It is unknown whether—." Presumably a first stage determines whether there are connections between John and gun. If none, subjects can say "unknown" rapidly. Otherwise a second stage carefully inspects the nature of the connection.

Some of the models that postulate a preliminary evaluation stage assume that the outcome of the evaluation not only determines whether or not to go on to do further processing, but also how much time should be devoted to this second stage. For example, R. Lachman, Lachman, and Thronesberry (1979) postulate metamemorial processes that regulate how long search continues for a specific piece of information: "Metamemory is accurate if it returns correct information about the contents in store. It is efficient if it appropriately controls search durations so that more time is allocated to seeking information actually present, less to information actually absent" (p. 543). Nelson et al. (1984) have found a positive relationship between the feeling of knowing and the amount of time elapsing before a memory search was terminated during recall.

Despite all the empirical support for the models described above, it is still uncertain whether initial evaluation of a question can influence strategy selection. This is because all of these models assume that if processing goes beyond a preliminary evaluation, there is only one possible strategy to be employed. That strategy is assumed to be a careful inspection of memory, in contrast to the more sloppy process used for the preliminary evaluation (e.g., Atkinson & Juola 1974, Glucksberg & McCloskey, 1981; R. Lachman et al., 1979; Smith et al., 1974). If a second strategy such as plausibility or inference is considered at all, it is assumed to always follow the direct retrieval strategy (e.g., J. L. Lachman & Lachman, 1980). The initial evaluation proposed in models such as Atkinson and Juola is, in reality, the first of a set of strategies to execute, and no selection is done at all. The goal of the following sections is to give support to the hypothesis that initial evaluation of a question can affect strategy selection.

Can We Rapidly Evaluate Our Ability to Answer Questions?

Past work on "feeling of knowing" has demonstrated that when people are unable to answer a question, they are nonetheless able to accurately assess the probability that they will be able to recognize the answer (e.g., Nelson et al., 1984; Nelson & Narens, 1980; Read & Bruce, 1982). Conceivably, the same type of mechanism that allows for these "feeling-of-knowing" judgments operates automatically even when people can answer the question. Given that question answering requires a strategy-selection stage prior to strategy execution, it seems reasonable that the mechanisms involved in "feeling of knowing" could be involved in assessing which strategy is preferable. For this to be a viable assumption, the "feeling of knowing" or initial evaluation would have to take substantially less time than the total time it takes to answer a question.

To see whether our "feeling-of-knowing" process could be involved in question answering, i.e., precede a strategy execution phase, Experiment 4 asks whether people can judge their ability to answer a question faster than they can actually answer it. If we do have a "feeling-of-knowing" process that enables us to select question answering strategies, then this assessment should operate faster as an explicit judgment than the task of actually retrieving the answer to the question, even though we have little practice at overtly assessing our "feeling of knowing."

EXPERIMENT 4

Game Show: Can People Estimate Answerability Faster Than They Can Answer?

To tap into this immediate, "feeling-of-knowing" process and compare it to question answering, a "game show" format was developed. Subjects see a question and rapidly decide whether they can answer it. If they respond "no" or wait too long after the question has appeared on the screen, they lose the opportunity to answer the question.

Method

Overview. Subjects were asked to read questions pertaining to world knowledge and then, depending on condition, either answer the question aloud or say whether or not they thought they would be able to come up with the answer. In both conditions, subjects were encouraged to respond as rapidly as they could. Those in the Estimate condition who said "yes" were then asked to come up with the answer to the question.

Subjects. Thirty-one Carnegie-Mellon undergraduates participated in the experiment to partially fulfill a course requirement. Fifteen were randomly assigned to the Answer condition and the other 16 to the Estimate condition.

Materials. Questions were constructed for four levels of difficulty: 20 extremely difficult or virtually impossible-to-answer questions (e.g., What is Menachim Begin's favorite dessert?); 20 difficult (e.g., Who was the first man to climb Mount Everest?); 21 questions of moderate difficulty (e.g., Where did the Greek gods live?); and 28 easy questions (e.g., How many tentacles does an octopus have?). The questions were all of approximately the same length; however, the primary concern was to compare performance between groups that both saw the same materials, so there was no attempt to ensure uniformity among sentence types. Assignment of questions to difficulty level was done on an intuitive basis, i.e., no norms were taken.

Procedure. The experiment was conducted on a terminal attached to a PDP/11 using the RSX-11 operating system. Attached to the terminal were a button box and a voice key with a microphone. Subjects were instructed about the task after their random assignment to one of the two groups. Those in the Answer condition were told to read each question on the computer screen and to say the answer into the microphone "as quickly as possible, without sacrificing accuracy." The answer triggered the voice key attached to the computer and the response latency was recorded. After responding, subjects typed their verbal response into the computer. Subjects were asked to say "don't know" as quickly as possible into the microphone when they did not know the answer.

Subjects in the Estimate condition were instructed to say "yes" or "no" into the microphone as quickly as possible after seeing the question. For "yes" responses, the subjects were then asked to type in the answer to the question.

To motivate fast responding, response times appeared on the computer screen for a few seconds after each trial. Subjects controlled the rate at which they saw the questions by pushing a start button to initiate the next trial. Subjects were also alerted to the problem of spurious triggerings of the voice key due to random noises such as coughs, and to the problem of responses spoken too softly to activate the voice key. On each trial, there was the opportunity to indicate that the response time was inaccurate due to early or late triggering of the voice key. The experimenter was present at all times to ensure that subjects typed in the verbal response they gave and to nullify trials where the voice key did not accurately record RT. The presentation order of the questions was randomly determined for each subject.

Results

A scoring program scored obviously correct responses. All other answers to a given question were presented to a rater to score. Because the computer presented all answers for a given question together, the rater had no idea whether a particular answer came from a subject ~~in the Es~~timate condition or the Answer condition.

Median response times were used to estimate each subject's performance in each condition. For each type of question, Table 5 gives the mean time to give the answer (Answer condition) or say "yes" (Estimate condition), the proportion of answers attempted (i.e., questions for which subjects did not say "don't know" in the Answer condition), the percentage of questions answered correctly, and the "accuracy" of the at-

TABLE 5
Time to Estimate or Attempt Answers, Proportion of Questions Attempted, Proportion Answered Correctly, and Accuracy of Attempts as a Function of Task and Question Difficulty, Experiment 4

Question difficulty	Time to attempt		Proportion attempted		Total correct		Accuracy of attempt	
	EST	ANS	EST	ANS	EST	ANS	EST	ANS
Easy	1.650	2.512	86.02	91.62	80.62	80.35	93.60	87.54
Moderate	1.828	2.654	67.46	72.29	59.48	54.97	87.98	76.08
Hard	1.728	2.388	42.38	47.09	34.13	28.42	81.44	59.07
Impossible	1.524	2.925	7.56	25.13	—	—	—	—
All attempted[a]	1.735	2.518	65.29	70.34	58.08	54.58	87.67	74.23

	Estimate	Answer
RT for answered correctly[a]	1.722	2.396
RT for answered incorrectly[a]	2.069	3.009
RT for not attempted[a]	1.851	3.197

[a] These means do not include the impossible items. The pattern is quite similar with their inclusion.

tempt (or calibration of "feeling of knowing"). Accuracy, viz., the ratio of proportion correctly answered over the proportion attempted, refers to how good the subject was at estimating what he or she knew. (For impossible questions, "can't say" was considered correct.) Table 5 also gives the mean RTs for questions correctly answered and incorrectly answered, and for questions for which subjects said "can't say." These latter numbers are not partitioned by difficulty of question, to save space. The analyses of variance used a 2 (task) × 3 (question difficulty—without impossibles) design for positive response times, regardless of whether or not they were correct.[7]

The most important result to note is that subjects asked to make estimates are over 25% faster than those asked to actually generate an answer. This difference is, of course, significant, $F(1,29) = 20.0$, $p < .001$. This is true whether one looks only at positive estimates or both positive and negative estimates, so it cannot be due to subjects merely saying a rapid "no" to anything they do not know very well in the Estimate condition. On the other hand, it is true that subjects in the Estimate condition attempt to answer fewer questions than do subjects in the Answer condition; however, this difference is not reliable ($F < 1.0$) and does not map onto fewer correct judgments in the Estimate condition. The groups do not differ in terms of the number of questions answered correctly ($F <$

[7] ANOVAs using only the correct answer times and those including impossibles look very similar and do not change any interpretations.

1.0). Since subjects in the Estimate condition attempt fewer questions, they have fewer erroneous attempts, making them significantly more accurate, where accuracy is defined as proportion correct of those attempted, $F(1,29) = 16.4$, $p < .001$. Clearly then, the speed advantage for the Estimate condition cannot be due to a speed/accuracy trade-off, where subjects are merely stopping the same process too soon.

There were significant effects due to difficulty of question, in terms of number of questions attempted, number answered correctly, and accuracy of attempts, $F(2,58) = 117.2$, 205.4, and 24.7, respectively, $p < .001$. Not surprisingly, subjects were less inclined to answer and less accurate with the more difficult questions. There was also an interaction of question difficulty with task, such that the accuracy advantage of the Estimate group over the Answer group was greater for more difficult question types, $F(2,58) = 4.1$, $p < .05$. Surprisingly, there was no difference in response times due to question difficulty.[8] On the other hand, in the next experiment, there is an effect of question difficulty on response times.

There is one other noteworthy result. The data were analyzed as a function of practice, partitioning the data into the first 25% of the experiment and the last 75%. Since the estimation task is not one that most subjects are used to performing, it seemed likely that any advantage of that condition would take time to develop. The advantage of the Estimate condition was 0.480 s during the first 25% of the experiment, but grew to 0.986 s in the last 75%.

Discussion

Although the data support the hypothesis that there is a mechanism that allows people to evaluate how much they know before they can actually answer a question, the results are open to other interpretations: In the Answer condition subjects have to articulate a longer response than subjects in the Estimate condition. There is evidence that subjects are faster to initiate an articulation of a short word (e.g., "yes") than a long word (e.g., "baseball") (e.g., Fowler, 1980; Klapp, 1974; Sternberg & Monsell, 1981). Those effects, however, tend to be on the order of 10 to 15 ms, while the effects reported here are on the order of 800 ms.

Given the sizable advantage of the Estimate condition, both in terms of

[8] It is unclear why the differences were not manifest in RTs. One explanation is that the difference among levels of difficulty was really probability of being able to answer the question, not difficulty of answering, per se. Consider, for example, one of the difficult questions, "Bowie Kuhn is (was) commissioner of what sport?" It is probably not difficult for those who know that answer (baseball). If a person knows the answer, it can be retrieved as fast as many questions considered much easier, e.g., "How many eggs are in a dozen?" The difference among the question types then would be seen only in measures such as the percentage of questions attempted.

response time and accuracy of estimation, it seemed worth demonstrating that the effect was not due to something as uninteresting as an advantage for getting to give the same "yes" and "no" responses on multiple trials. Therefore, in order to control for any advantage due to binary responding, per se, Experiment 5 required subjects in both groups to make binary decisions prior to giving the answer.

EXPERIMENT 5

Binary Responding for Estimate vs Answer

This experiment was similar to Experiment 4 with several notable exceptions: Subjects in both groups first pushed one of two buttons. If they pushed the button indicating that they had the answer ready to give (Answer condition) or thought they could answer the question (Estimate condition), then they went on and said the answer into a microphone that recorded their latency to generate the answer. Two response times were collected for those questions that had an affirmative response, namely time for the positive response and time to articulate the answer.

Getting subjects in the two conditions to treat the two tasks differently was not trivial since the logical structure of the tasks was identical. The answer group was penalized if they could not come up with the answer shortly after pressing the button. Further, in the Answer condition, the question was removed from the screen after pressing the button, since subjects were already supposed to have the answer in mind; however, in the Estimate condition, the question remained on the screen until the subject said the answer into the microphone. They were given unlimited time to give the answer after estimating that it was answerable.

Method

Materials and design. Both the questions used and the design of the experiment were identical to Experiment 4. The only difference was in the collection of two latencies per trial, rather than one.

Procedure. The experiment was conducted on an IBM personal computer, to which a button box, a microphone, and voice key were connected. Subjects were told that the experiment was similar to a television game show; they would accumulate points for their answers, which were redeemable for cash at the end of the experiment.

Subjects assigned to the Answer condition were instructed to press the green (right-hand) button as soon as they were sure they had the answer to a question and were ready to say it, but to press the red button (left-hand) if they were sure they did not know the answer. They were told to indicate their response only after searching through memory and either finding or failing to find the answer. Those in the Estimate condition were instructed to press the green button as soon as they thought they probably would be able to find the answer in memory, and to press the red button as soon as they thought they probably would not be able to find the answer.

Both groups were given points for fast responding to the first button press; however, in the answer group, subjects had to speak the answer into the microphone within 1 s of pressing the button or else they lost points. The points awarded for each answer depended on button press response time. Additionally, for affirmative responses, points depended on accuracy of answer and whether or not the verbal response was begun within the 1-s time limit (in the Answer condition). The scoring method was explained to both groups, and subjects were told how many points they had accumulated after each trial. They also earned points for negative responses. Because it was easier to amass points in the estimate group, the conversion of points into money was different for the two groups.[9]

Following an affirmative response, the computer screen prompted the subject to say the answer into the microphone. If a subject in the Answer condition failed to give the answer within 1 s, the computer then prompted the subject to still attempt to give the answer (but implied that the response was late). After a verbal response was given, the correct answer was displayed on the screen and the experimenter scored the response. Questions were presented in random order.

Subjects. Thirty-three undergraduates enrolled in their first psychology class participated to partially fulfill a course requirement. No subject had participated in a previous version of this experiment. Sixteen subjects were randomly assigned to the Estimate condition and 17 to the Answer condition. In addition to receiving course credit, they received nominal payment for performance in the task: 2.5 mils/point in the estimate group and 3.125 mils/point in the answer group, which averaged about $0.50 in bonus payment.

Results

Table 6 presents the data in a format similar to Table 5. The estimation or attempt times for questions that were subsequently answered correctly (RT1) are given for each question type instead of showing all positive response times. (The difference between the two measures is very slight.) The times are also given for the correct articulation of the answers, (RT2), and the sum of these two response times (RT1 + RT2).

The ANOVAs used the same factors as Experiment 4, and used correct RTs for Phase 2 and the sum of these two times. The percentage of questions attempted did not differ for the two instructional groups, although it did differ as a function of the difficulty of the questions, $F(2,62) = 153$, $p < .001$. Percentage correctly answered and accuracy did not differ across the two groups, but again, did differ with difficulty of question, $F(2,62) = 180.85$ and 18.44, respectively, $p < .001$.

Of greater interest, time to estimate that a question could be answered was significantly faster than time to indicate an answer was "in mind," $F(1,31) = 4.78$, $p < .05$. Response times for the first phase also differed significantly as a function of question difficulty (unlike Experiment 4), $F(2,62) = 4.24$, $p < .05$, such that for easier questions, subjects were faster at being ready to give the answer or to estimate they could answer them.

[9] The payoff formula was: $3.8/(0.15 \times rt + 0.4) - 3.7$, rounded to the nearest half point. In the estimate group, subjects were given an additional 3 points for a correct response and lost 3 points for an incorrect response. In the answer group, they received the same payoff for accurate and inaccurate responses so long as they were given in less than 1 s after the button press; if it took longer than that to speak the answer, subjects lost 4 points for a correct answer and 6 for an incorrect answer.

TABLE 6

Mean Proportion Attempted, Correct, Accuracy and Response Times to Attempt Correct Answers, Give Answers, as a Function of Task and Question Difficulty in Experiment 5

Question difficulty	RT1: Time to attempt		Proportion attempted		Total correct		Accuracy of attempt	
	EST	ANS	EST	ANS	EST	ANS	EST	ANS
Easy	1.385	1.580	84.13	85.85	78.70	79.89	93.27	95.99
Moderate	1.447	1.671	68.69	68.94	60.67	59.83	88.43	88.40
Hard	1.427	1.775	40.47	37.51	32.69	29.96	80.98	80.34
Impossible	—[a]	—	6.68	5.00	—	—	—	—
All attempted[b]	1.420	1.675	64.43	64.10	57.35	56.56	87.56	88.24

	RT2		RT1 + RT2	
	EST	ANS[c]	EST	ANS[c]
Easy	0.590	0.304	1.975	1.884
Moderate	0.595	0.274	2.042	1.945
Hard	0.541	0.317	1.968	2.092
Impossible	—	—	—	—
All attempted[b]	0.575	0.298	1.995	1.974

[a] —, undefined.

[b] These means do not include the impossible items.

[c] RT2 in Answer condition does not include late responses (less than 1% of data).

Time to generate the answer also differed significantly as a function of task instructions, $F(1,31) = 9.18$, $p < .01$, but in the opposite direction. Subjects in the Answer condition were faster than the Estimate condition, as they should be, to give the answer. Question difficulty did not affect time to give the answer, nor was the interaction with task significant. As the model would predict, the sum of the two response times did not differ significantly, $F(1,31) = 0.02$, for the two tasks, even though the time for each part differed significantly (but in opposite directions). The reason the model would predict that the sums would be roughly equivalent is that the estimate phase RT should be a subset of the answer task's first phase, and the processing that the answer task did during the first phase, namely finding the answer in memory, should be included in the RT for the second phase for the estimate group.[10]

Discussion

Experiment 5 replicates the findings of Experiment 4, that subjects can estimate that they can answer a question significantly faster (without sacrificing accuracy) than they can actually find the answer. That is, subjects

[10] The reason why the RTs for the first phase are so much longer than the second phase is that the first-phase RTs include the reading time of the question.

are at least as accurate at estimating answerability as they are at attempting the answer. It is unlikely that this advantage is due to something trivial such as difficulty in articulating the response, since both groups made a binary decision followed by the complete answer. Taken together, these data support the proposal that we have the capability to assess our memories before we do a careful search of memory.

The research reviewed at the beginning of the paper and the first set of experiments argued strongly that strategy selection is part of question answering. These last two experiments have shown that a sentence can receive an initial evaluation quickly enough to make it reasonable that intrinsic variables can also influence that selection. Below I outline the kinds of mechanisms and cognitive factors that might be involved in the initial evaluation process.

Mechanism for Evaluating "Feeling of Knowing"

The process involved in the initial evaluation of a question might include (1) determining how recently the terms in the question have been encountered and (2) measuring the extent of knowledge stored in memory relevant to the question. These processes are assumed to operate in the context of a semantic network in which concepts in memory can become "active" from the terms in the test question. Recency (to be referred to as familiarity) and extent of related information are measured in terms of activation.

Familiarity. Although there have not been models concerned with how feeling of knowing might affect strategy selection, there are theories concerned with how familiarity is determined. For the most part, these theories (e.g., Jacoby & Dallas, 1981; Hasher & Zacks, 1979; Hintzman, Nozawa, & Irmscher, 1982; Mandler, 1980; Zacks, Hasher, & Sanft, 1982) have postulated two separate mechanisms for judging familiarity. Although these ideas have been applied mostly to tasks concerned with recognition or frequency judgments, they can be incorporated into a framework which is concerned with more complicated types of memory queries.

Mandler (1980) has argued for two separate types of recognition processes, one that measures familiarity or occurrence information, and the second that is a much slower, more careful retrieval mechanism or search. He suggests that the first type of process is affected by shifts in modality (e.g., auditory during study, but written at test) and that this familiarity/occurrence information decays faster than the propositional/symbolic information; however, the familiarity process also can execute faster during recognition. This proposal of an automatic process that recognizes familiar traces is similar to the proposal of Hasher and Zacks (1979) that there is an automatic mechanism that keeps track of frequency information. Hasher and Zacks, like Mandler, also postulate a more "controlled" (nonautomatic) memory mechanism. They found that many variables which affect recall performance do not affect the frequency judgment performance, i.e., the latter process does not degrade with increased processing loads, age, etc. Hintzman et al. (1982) also have data consistent with these theories. Jacoby and Dallas (1981) make similar proposals as well; specifically, they postulate two memory types, an autobiographical form of memory and a "less aware" form of "perceptual learning." They note that levels of processing (see Craik & Lockhart, 1972) affect recognition memory but not perceptual recognition. Perceptual recognition might be thought of as a physical match, and is like the mechanism that keeps track of frequency information for Hasher and Zacks (1979) and Hintzman et al. (1982), and is also like the fast, recognition-memory mechanism of Mandler.

In the present framework, determining the recency of exposure to a concept in the question is measured by how active it is relative to its base activation level.[11] So, for example, if a story mentioned certain words often and some of those words were contained in a test statement, the feeling-of-knowing mechanism would probably register high familiarity. Consistent with this view, Koriat and Lieblich (1977) found that feeling of knowing for a nonrecalled item is increased by repeating the question more than once or by adding redundant terms to the test probe.

Related knowledge in memory. In addition to the fast process of determining "raw familiarity," this initial evaluation also measures the "relatedness" of the concepts in the question through the interconnections in memory. The proposal that relatedness affects feeling of knowing has also been discussed by Nelson et al. (1984). They consider a number of factors that might affect recall failure and feeling of knowing. They suggest that "related episodic information" may influence a feeling of knowing judgment. There is also evidence that relatedness affects decision times. Rips, Shoben, and Smith (1973) postulate differences in feature overlap to explain the faster categorization times for dominant instances. Some research of my own (Reder & Anderson, 1980; Reder & Ross, 1983) also suggests that subjects can use a relatedness judgment to bypass retrieval of a specific fact from memory. "Relatedness" (for initial evaluation) is defined as the degree to which words in a question cause activation to intersect in memory. The more intersections detected

[11] A common word like *table* would have to be much more active than a rare word like *hippopotamus* for the feeling-of-knowing process to recognize it as having been seen recently. See, for example, Just and Carpenter (1980) or McClelland and Rumelhart, (1981) for a fuller discussion of similar assumptions.

in memory as a result of a query, the more potentially relevant information is available for question answering.

These two feeling-of-knowing processes, familiarity detection and intersection detection, go on in parallel. It is a useful heuristic to assume that when familiarity is high, the statement was seen recently and should be relatively accessible. Direct retrieval is a faster and easier strategy than judging plausibility when the specific fact that must be found is relatively accessible. Therefore, when familiarity is high, direct retrieval should be the preferred strategy.

When the process that detects intersection of activation determines that there is a lot or a moderate amount of potentially relevant information, there is a bias to use the plausibility strategy. When both biases exist, then the bias to use plausibility is superseded by the bias to use direct retrieval. This is because plausibility always has a longer computation stage, and if the memory search is relatively easy, plausibility does not have the compensating search time advantage to make it the preferred strategy (Reder, 1982). Questions that produce little activation are immediately "recognized" as unanswerable (e.g., what is the rate of mitosis in paramecia?).[12]

This next experiment tests a few of the implications of this initial evaluation based on intrinsic features. It is designed to see whether our "feeling of knowing" is really based on things like familiarity (recency of exposure) and number of intersections in memory, even when it turns out that these features have no predictive validity as to the subject's knowing the answer.

EXPERIMENT 6

Can Our Estimation Process Be Subverted by Spurious Familiarity?

Like Experiments 4 and 5, in this experiment half of the subjects were asked to estimate the answerability of questions and the other half to answer them directly. Before subjects began the question-answering or estimation phase, they were asked to rate the frequency of occurrence of some terms or pairs of terms. These terms were selected from a random third of the questions to be estimated and/or answered. Rating terms that would be seen as part of a question was the "priming" manipulation. Half of the subjects were asked to rate pairs of terms on conjoint frequency (where both terms were taken from the same question); the other

[12] Some questions, such as, "What is Dickens' phone number?" may seem to be rejected from this initial evaluation; however, it is more likely that the decision is made to use a particular question-answering strategy, and then it is determined that the question is unanswerable.

half rated terms in isolation. For both types of rating groups, the same terms from a question were rated if a question was to be primed. For this reason, subjects who rated pairs had exactly half as many rating trials.

In the past, subjects asked to estimate whether they could answer a question were more accurate than subjects actually asked to answer them. The prediction is that by priming words from a question, subjects would be "thrown off" in terms of using the mechanisms they normally use to judge answerability. The estimate group should estimate that they can answer more primed questions than unprimed, at least for those questions that are difficult, i.e., that they would not otherwise judge as answerable. Question difficulty was varied systematically so that this prediction could be tested. The answer group was not expected to give more answers to the primed questions; however, they were expected to take longer to say that they could not answer a question (i.e., search longer for the answer before giving up) if it had been primed. Also of interest was whether the effect of priming differs depending on whether the terms were rated together or separately.

Method

Materials. Questions were selected from a set that has been normed for answerability (Nelson & Narens, 1980). Three levels of difficulty were used: 21 questions from the most difficult of Nelson and Naren's question set were selected (mean recall = 6.5%); 21 from midrange (31.8%); and 21 from the easiest third (71.5%). For each question, two terms that seemed (a) least common and (b) most "central" to the question were selected to be the candidate priming words. For example, for the question "What term in golf refers to a score of 1 under par on a particular hole?" the priming terms were "golf" and "par."[13] Candidates were needed for all questions, since those selected for priming within each level of difficulty were randomly determined for each subject. In addition to these 63 questions, 15 easy practice questions were used.

Design. There were two between-subject factors: task group (estimate vs answer), and type of priming or rating task (rating individual term vs rating term pairs). There were also two within-subject factors: question difficulty (easy vs moderate vs difficult) and whether the question was primed by the rating task or not (primed vs unprimed). Half as many questions were primed as unprimed, because subjects might have become suspicious of the priming manipulation and attempted to alter their "feeling-of-knowing" strategy if too many questions were primed. Each level of difficulty had 7 primed and 14 unprimed questions.

Procedure. Subjects were seated in front of an IBM PC and randomly assigned to one of four conditions. (They were unaware initially as to whether they would be making estimates or directly answering questions.) Before the question phase began, subjects were told to rate the terms that would be displayed on the screen. Subjects were asked to rate, on a five-point scale, how often a term was encountered during reading or listening, or how often the pair of terms was encountered together, i.e., conjoint frequency. The order of presentation of the terms or pairs was randomly determined for each subject with two constraints:

[13] The terms were sometimes longer than one word, e.g., "Lady Godiva," "Howdy Doody."

TABLE 7
Proportion of Questions Attempted, Accuracy of Attempt, Time to Attempt Answer, or Time to Say "Don't Know" as a Function of Task, Question Difficulty, and Priming

	% Attempted		Acc of attempt		RT:Attempt		RT:Not attempt	
	Unpr	Primed	Unpr	Primed	Unpr	Primed	Unpr	Primed
				Estimate				
Easy	.82	.76	.96	.91	2.41	2.32	3.80	3.45
Mod.	.50	.52	.79	.75	2.79	2.58	3.05	3.34
Hard	.24	.31	.56	.53	2.72	2.70	2.80	2.75
				Answer				
Easy	.77	.81	.93	.89	3.42	3.64	7.56	6.76
Mod.	.54	.48	.67	.71	4.58	4.67	6.04	8.30
Hard	.39	.38	.36	.42	5.10	5.47	4.59	5.51

Terms from the same sentence could not be presented sequentially in the "single" conditions, and practice items always were rated first.

Following the rating task, subjects were instructed about the question-answering phase, which was similar to Experiment 4. The answer group spoke their answers directly into the microphone, and the estimate group said "yes" or "no" orally before giving an answer.

Subjects. There were 76 subjects: 18 in the Estimate-single group, 20 in the Estimate-pair group, and 19 in each Answer condition. Forty-seven subjects were paid $4.50 for participating in this and one other experiment. The others were given course credit. The paid subjects were either students or staff from Carnegie-Mellon, while those receiving credit were students participating in their first psychology course.

Results

Table 7 is organized so that the data from subjects in the estimate groups are presented on top, and the data from subjects in the answer groups are given in the lower panel. The data are given on separate rows for the three levels of question difficulty. Each row gives the proportion of questions attempted, the accuracy of the attempts (percentage correct divided by percentage attempted), time to attempt an answer (say "yes" in the Estimate condition), and the time to say "don't know" (or "no"), for both unprimed and primed questions. The data are collapsed over the type of rating task (pairs vs singles), since the patterns are very similar for the two priming conditions and that variable did not interact with any other.[14]

[14] Subjects were significantly faster and more likely to attempt answers in the paired conditions, for both the estimate and the answer task, although much more so for the estimate task. The faster judgments did not interact with other variables. Although it is possible to construct explanations for the modest differences across conditions, it is more likely that any differences are due to subject effects, and dwelling on these slight differences would confuse the picture.

Analyses of variance were performed for each of the measures listed in Table 7, using the factors listed there and rating condition. The median of a subject's times in a condition was used in the analyses. For purposes of the ANOVAs, missing cells were estimated by using the grand mean of each group plus the individual's subject effect, plus the relevant condition's effect.[15]

Effect of priming on proportion of questions attempted. The proportion of questions attempted varied as a function of difficulty, and also, difficulty interacted with task (answer vs estimate), $F(2,144) = 3.08$ and 6.01, respectively, $p < .01$. For both tasks fewer hard questions were attempted, but the drop-off from easy to hard was more precipitous in the estimate task (replicating past results). Of more interest, there was a significant interaction of task with question difficulty and the priming variable, $F(2,144) = 4.52$, $p < .01$, one tailed; $p = .0125$, two tailed). In the estimate task, subjects estimate 6% fewer primed questions than unprimed questions in the easy task, 3% more primed questions of the moderately difficult, and 7% more of the hard ones. This represents a significant interaction of question difficulty with priming, $F(2,72) = 3.61$, $p < .05$, for the estimate subjects.

There was no systematic effect of priming for proportion of questions attempted in the answer task. The interaction of question difficulty with priming was not significant in the answer task, $p > .10$. No interaction was expected there either, since people could not just use their "feeling of knowing" in order to answer. The effect of priming was not consistent across the two rating tasks in the answer task, while the pattern of priming effects for levels of difficulty was the same for both rating tasks in the estimate task.

Effect of priming on time to respond. The time to attempt an answer was, of course, longer for difficult questions, $F(2,144) = 17.54$, $p < .01$. There was also a significant interaction of difficulty with type of task, $F(2,144) = 6.32$, $p < .01$. The slowdown in RT with more difficult questions was greater in the Answer condition because subjects had to actually find the difficult answers, while in the Estimate condition, a "feeling of knowing" may come almost as quickly for difficult questions.

Priming had opposite effects on the two tasks: Subjects were over

[15] The means in the Tables reflect the obtained scores without the estimates for missing or undefined observations. The F statistics vary depending on the estimation procedure used for the missing observations. The Fs in question are the response times. For example, if a subject did not attempt any difficult primed questions, then an estimate of the time for that subject to attempt a difficult, primed question was required. The Fs reported were the most conservative, i.e., gave the smallest Fs, of the various estimation procedures. Proportion attempted, on the other hand, was not affected since there were no missing observations. Therefore, those statistics are not subject to interpretation.

0.100 s faster to estimate that they could answer questions if the questions had been primed. On the other hand, subjects were 0.200 s slower to give an answer if the question had been primed.[16] Presumably the priming facilitated the "feeling of knowing" for both groups. For the estimate task, that was all that was needed, and a decision could be made sooner. For the answer group, however, this "feeling of knowing" led them to look longer for answers than they would have otherwise.

Effects of priming on times to not attempt an answer. The result that subjects were slowed down by priming in the answer task is also reflected in the times for those questions that were not attempted. There is a marginally significant interaction of priming and task, $F(1,72) = 3.3, p < .10$, reflecting the fact that subjects were 0.800 s slower to say they could not answer a question if it had been primed but were 0.035 s faster for primed questions if they only had to estimate that they could answer them. There were significant interactions of Difficulty × Task, $F(2,144) = 7.0, p < .01$, and Difficulty × Priming × Task, $F(2,144) = 3.1, p < .05$. The rating task gave subjects a false "feeling of knowing" in the answer task too. For the more difficult primed statements, they therefore tried longer to come up with an answer before realizing that they could not.

Effects of priming on accuracy of attempts to answer. The false "feeling of knowing" can also be seen in the effect of priming on subject's accuracy of "feeling of knowing." In the Estimate condition subjects are consistently better calibrated when the questions have not been primed, although this trend is not significant. One would not expect the pattern in the Answer condition and it did not occur. If anything, the opposite result occurred, viz., that subjects in the Answer condition are more accurate when they answer if the question has been primed. The fact that priming hurts estimation accuracy is converging evidence for the view that people can use recency of exposure to influence their feeling of knowing.

It is worth knowing that although priming makes people more inclined to think they can answer a question, difficulty of questions has a greater impact: Attempts drop from about 80 to 30%, and accuracy of attempts drops from about 90% to only a little above 50%. This suggests that variables other than recency of exposure are having greater impact on "feeling of knowing." Indeed, Reder and Fabri (1982) have shown that people's rating of knowledge of topic is a strong predictor of whether or not they estimate that they can answer a specific question—this perceived knowledge of a topic predicts estimation behavior. It more strongly predicts attempts to answer in the estimate task than it does in

[16] Although this interaction was not reliable, it was reliable with other procedures used for estimating missing data.

the answer task (and accuracy of attempts is better in the estimate task, too).

Finally, it is also noteworthy that the general level of accuracy (or calibration) in the estimate tasks in this experiment as in Experiments 4 and 5 was much better than that reported elsewhere (e.g., Nelson, Leonesio, Landwehr, & Narens, 1986). In my estimate tasks, accuracy was often over 80% while their correlations were closer to 30%. I suspect that the reason for this is that others' correlations were based solely on those items that subjects could not originally answer, while my measures are taken from a greater proportion of trials, namely all trials where subjects attempted to answer. Correlations of restricted ranges tend to be lower.[17]

Discussion

The results of Experiment 6 support the idea that recent exposure to concepts affects a person's "feeling of knowing." The results also support the idea that there is a separate process for initial evaluation, since the manipulation of priming had complementary effects for subjects asked to answer directly as compared to those asked to estimate whether or not they could answer: Proportion attempted was affected by priming only in the estimate task, since subjects in the Answer condition had to come up with the answer; on the other hand, time to say "don't know" was only affected by priming in the answer tasks.

One result in the data to be explained is why subjects were less inclined to estimate that they knew an answer to a question for primed questions that were easy to answer. One explanation is that subjects were aware of the priming manipulation and its distorting influences and were trying to counteract it. If subjects were aware that they were more inclined to positively estimate when the items seemed familiar from rating, and that their accuracy suffered as a result, they might try to counteract the manipulation. If so, whenever they recognized that the terms in the probe had been previously rated, they might raise their criterion for "feeling of knowing."

The raising of the criterion would not have the same effect for all levels of question difficulty. This is because priming does not have the same effect for easy as it has hard questions. Hard questions are influenced much more by priming, since easy ones would be attempted anyway. This correction procedure (raising the criterion) does not completely counteract the priming manipulation, and therefore, hard questions still

[17] I did not measure accuracy of estimation for those subjects did not attempt; however, I doubt that that accounts for the difference in calibration. In the answer condition, subjects attempted more, but rarely were successful. This suggests that those same items in the Estimate condition would also be wrongly answered.

show a bias in feeling of knowing; however, since priming was never needed for easy questions, this correction lowers estimates for primed, easy questions.

This explanation is consistent with the theory, discussed earlier, that people are sensitive to contextual variables in the testing situation, e.g., notice the effects of the priming manipulation. We know that people are capable of rapidly altering their strategy. It seems that subjects are trying to alter the strategy that they would otherwise select, on the basis of such strategic decisions as "this probe contains terms that were in the rating task—therefore I should underestimate how easy it would be to answer."

GENERAL DISCUSSION: PUTTING THE TWO STRATEGY-SELECTION PROCESSES TOGETHER

This paper has argued that there exists a strategy selection stage and suggested two types of processes that are involved in the selection: One type is strategic processes that evaluate situational or contextual information, and the other type is less conscious processes that quickly evaluate how familiar the question seems. Experiments 1, 2, and 3 showed that we are quite sensitive to extrinsic factors when selecting a strategy. Experiments 4 and 5 showed that we "evaluate" sentences fast enough for initial evaluation of a question to be part of the strategy selection process. Experiment 6 indicated that recency of exposure to words influences our "feeling-of-knowing" process, postulated to be involved in strategy selection.

Both types of processes generate a bias for the strategy to be selected. If the two biases are in conflict, the stronger bias prevails. An example of the blending of these two processes comes from the work of Gentner and Collins (1981). They found that people are more likely to decide that an assertion is implausible as the assertion's importance and their own expertise relevant to the assertion increase. This is the "I would know this fact if it were true" phenomenon. If a person knows a lot about an area (strategic information), and there is no convergence of activation from the concepts in the assertion (initial evaluation), a person may be more willing to make a quick "no" response. Alternatively, if there is some convergence of activation, then knowing more about an area might lead one to spend more time searching for the information (Collins, personal communication).

How Does the Framework Relate to Current Views of Question Answering?

One of the major influences on theories of question answering comes from the Cognitive Science Group at Yale (e.g., Kolodner, 1983, 1984;

Reiser et al., 1985; Reiser, Black, & Kalamarides 1986; Schank, 1982). Schank's (1972) original view was that all inferences are computed "on-line" during comprehension. In this way, when a question is asked that does not tap information explicitly presented, the answerer can look at the memory representation of the input and directly retrieve the information, since it was already inferred (e.g., Schank, 1972, 1975; Schank & Abelson, 1977). Lehnert's (1978) work on question answering involved inferential mechanisms at the time of test; however, these were not to compute the answer to the question, per se. Rather, the inference was to determine the true intention of the question asker (e.g., some questions are really requests, as in "it is chilly in here, don't you think?" means "please close the window") or to figure out what level of answer is appropriate (e.g., would a yes or no answer be enough). In her model, the answering mechanism can still be thought of as direct retrieval once the question is properly determined. Graesser (1981) also seems to believe that large quantities of inferences are constructed during comprehension and that these will be used for question answering if available. In other words, none of these theories consider that plausible reasoning could be a preferred question-answering strategy.

Singer and Ferreira (1983) do not assume that "all" plausible inferences are made in advance. Nonetheless, they, too, believe that if an inference is made "on-line," that it will be retrieved. Neither a stored inference nor a presented statement would ever be verified by inference at time of test if it could be found in memory. Test statements that were not inferred during comprehension would have to be computed, but that is the less preferred strategy. The model proposed here is different from both of these points of view. Regardless of whether or not the inference was computed "on-line" during comprehension, there is a very strong possibility that a person will not bother to try to find it. Instead, people will often just try to compute or recompute whether or not an assertion seems plausible.

In addition, Kolodner (1980) has looked at the retrieval from memory of personal or "episodic" (Tulving, 1972) information. Reiser (1983) has also looked at retrieval of personal memories and has developed a model similar in spirit to Kolodner's. He has also gathered empirical support for his model (Reiser, 1983; Reiser et al., 1985). The type of memory queries that their theories address concern events, e.g., "did you ever go swimming in a river in Ohio?" In their framework inferential mechanisms are required to answer questions; however, in this case, the inferencing must be done to extract the relevant search contexts and infer whether an item not found during initial search can be found with further search. Their view is that one memory retrieval can provide cues for a subsequent retrieval. This is similar to ideas described by Norman and Bobrow (1979), Williams and Hollan (1981), and Williams and Santos-Williams (1980). In their research, too, memory retrieval was viewed as a recursive, reconstructive, problem-solving process. That is, search is a cycle of specification, matching, and evaluation that continually refines the descriptions of the items sought in light of the evaluation. Search for an item retrieves partial information, which is used to build a more complete specification of the target to guide further searches.[18]

The successive refinement strategy for finding information in memory proposed by Kolodner and Reiser could be thought of as yet another mechanism that can be used to answer questions. If this type of strategy were selected for answering the question, the answerer would have to predict a plausible memory location for an event that might have the relevant target features. When the features sought are not found there, a new location is tried or an assessment is made of the probability of finding the required information. This type of strategy was not appropriate in the experiments I reported, because the information sought was (a) not autobiographical and (b) relatively recent. Indeed, Kolodner and Reiser also distinguished "Question Answering" from their much harder memory search, where the former is considered appropriate to answering questions from stories. For example, Kolodner states that search for relevant contexts is not needed for question answering of stories read recently.

The framework developed in this paper can be expanded to include this other type of question-answering task. During the first stage that measures "feeling of knowing," a decision that some relevant information can probably be found would bias Stage 2 against using the direct retrieval strategy, because the terms in a probe of this kind would not register as having been seen recently. Because the question deals with a specific episode or event, an evaluation would be made that the "successive-refinement" search strategy would be needed. Before going on to search for the relevant information, an inference would be made as to the appropriate first context to search.

How Does the Current Framework Relate to Other Areas of Cognitive Processing?

One idea developed in this paper is that we have a decision mechanism to select the dominant strategy to answer a question. This idea is in contrast to some of the extant models of cognition, e.g., Anderson's (1983) ACT* model, where productions can not be given probabilities of firing. Productions are either part of the goal set or they are not. It is not obvious how the production sets or goals would be easily reordered in dom-

[18] The proposal that one retrieval can facilitate another retrieval has also appeared in models of free and cued recall, e.g., Shiffrin (1970); Raaijmakers and Shiffrin, (1981).

inance, based on advice given from trial to trial. It is also not obvious how these ideas would be implemented by theories which propose that processing is implemented in massively parallel architectures (e.g., McClelland, Rumelhart, & Hinton (1986); Fahlmann, Hinton, & Sejnowski, 1983). As yet, these models lack any obvious mechanisms for allowing a person to prefer one process to another, varying from trial to trial.

The mechanisms for deciding which strategy to select, or how to allocate attention among competing question-answering strategies, may be quite similar to those mechanisms that allow a person to allocate attention in a dual-task situation. There is a large body of literature concerned with allocation of resourcqs when attempting to perform multiple tasks. When driving a car, the amount of resources devoted to driving as opposed to a second task such as conversing with a passenger can shift as the attention demands of the driving task change.

Just as "feeling of knowing" can affect what strategy is selected in question answering, "feeling of competence" could affect how much attention is devoted to one task as opposed to another. The controlled, strategic processes that are influenced by external factors in question answering could also monitor the feedback from the environment in the dual-task setting to see whether the tasks are being performed well enough (or one too well, and the other not well enough). Advice such as "watch the road" or "listen to this argument for why my theory is better" may cause the allocation of resources to shift, just as the advice to "try to find the fact in memory" or "try to infer the answer" can affect question-answering behavior.

A considerable portion of the research on attention concerns allocation of resources in a dual-task situation (e.g., Gopher & North, 1974, 1977; Kinchla, 1980; LaBerge, 1975; McLeod, 1977; Moray, 1967; Moray & Fitter, 1973; Navon & Gopher, 1979, 1980; Norman & Bobrow, 1979; Schneider & Shiffrin, 1977; Shaw, 1980; Sperling & Melchner, 1979; Wickens, 1980). Some of this research is concerned with the issue of whether or not the processes for dual tasks operate in parallel or whether they occur sequentially, whether the subject is strategic in resource allocation between the two tasks, and whether time-sharing or resource allocation abilities are learned and can be improved (e.g., Damos & Wickens, 1980; Fisher, 1975a, 1975b, 1977, 1980; Lane, 1980; Schweickert, 1978, 1980, 1983; Shaw, 1980).

Work by Posner, Nissen, and Ogden (1978) gave impetus to the idea that we can allocate attention differentially by using strategic information. They looked at how response times varied to the onset of a light as a function of the validity of a directional cue. Subjects performed faster with a valid cue (correct advice as to which position would light up) than when not given a cue, and performed slower with an invalid cue.

In this literature, the strategic components that allocate resources are called "time-sharing" skills. Lane (1980) and Damos and Wickens (1980) have looked at the acquisition of time-sharing skills. Lane found that the difference between central and incidental task performance increases with age, and that the correlation between the two tasks decreases with age. However, unlike the common assumption that such developmental differences are due to acquired ability to process information selectively, they may be due to capacity trade-offs, i.e., devoting most of the processing resources to one of the dual tasks. Damos and Wickens found that time-sharing skills are learned with practice and can transfer to new dual-task combinations. Navon and Gopher (1979) also found that with extensive practice, subjects seemed able to achieve higher levels of performance on both tasks. One possibility they consider is that the processes involved in the allocation of resources get better with practice.

EXTENSIONS

Although the strategies involved in other cognitive tasks are undoubtedly different from those for question answering, it is interesting to conjecture on the similarity of the factors affecting strategy selection. Consider, for example, mathematics. We can first assess our familiarity with a problem. Certain classes of problems we recognize that we can work out without consulting a math book, while others require looking up the formula for solution. Even within multiplication we may assess how familiar a problem is before selecting a procedure for solution. Common but complex problems such as 12 × 12 are recognized as directly stored, but 16 × 12 might be evaluated as one that should be broken down into subcomponents to solve. Siegler and Shrager (1984) have evidence that children do similar things for arithmetic, although they apparently always try direct retrieval before computing the answer.

In addition to familiarity evaluation, it is probably also the case that in many domains prior history of success with a strategy influences selection of the strategy. Again, using mathematics as an example, consider the task of solving integrals. If one integration technique has worked for many problems, one is likely to try it again unless the initial evaluation suggests some other procedure. The Einstellung effect in solving water jug problems (Luchins, 1942) is an example of where prior history of success with a (solution) strategy adversely affects strategy selection. The Einstellung effect is eliminated for approximately 50% of the subjects when they are simply instructed "Don't be blind." This is understandable since "prior history of success with a strategy" is part of a selection mechanism under conscious control.

For almost any complex cognitive task there are multiple strategies that can be used for solving it, typically one of which is more appropriate than another. As people become proficient in performing the task, they also become relatively proficient in selecting the best strategy to use for a particular instance. This paper has provided a framework for understanding the role of strategy selection in question answering and has suggested what variables affect the selection process.

REFERENCES

Anderson, J. R. (1972). FRAN: A simulation model of free recall. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 5). New York: Academic Press.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard Univ. Press.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Hemisphere.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 1). San Francisco: W. H. Freeman.

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: Advances in research and theory* (Vol. 2). New York: Academic Press.

Bower, G. H. (1972). Mental imagery and associative learning. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley.

Camp, C. J., Lachman, J. L., & Lachman, R. (1980). Evidence for direct-access and inferential retrieval in question-answering. *Journal of Verbal Learning and Verbal Behavior*, 19, 583–596.

Collins, A. (1978a). Fragments of a theory of human plausible reasoning. In D. L. Waltz (Ed.), *Theoretical issues in natural language processing*. Urbana–Champaign, IL: University of Illinois.

Collins, A. (1978b). *Studies of plausible reasoning: Vol. 1. Human plausible reasoning*. (BBN Rep. No. 3810). Bolt Beranek & Newman.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240–247.

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

Damos, D. L., & Wickens, C. D. (1980). The identification and transfer of time-sharing skills. *Acta Psychologica*, 46(1), 15–39.

Erikson, T. A., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbul Behavior*, 20, 540–552.

Fahlmann, S. E., Hinton, G. E., & Sejnowski, T. J. (1983). Massively parallel architectures for AI: NETL, Thistle, and Boltzmann machines. In *Proceedings of the National Conference on Artificial Intelligence AAAI-83*. Washington, DC.

Fisher, S. (1975a). The microstructure of dual-task interaction 1. The patterning of main-task responses within secondary-task intervals. *Perception*, 4, 267–290.

Fisher, S. (1975b). The microstructure of dual-task interaction 2. The effect of task instructions of a model of attention switching. *Perception*, 4, 459–474.

Fisher, S. (1977). The microstructure of dual-task interaction 3. Incompatibility and attention switching. *Perception*, 6, 467–477.

Fisher, S. (1980). The microstructure of dual-task interaction 4. Sleep deprivation and the control of attention. *Perception*, 9, 327–337.

Fowler, C. A. (1980). Coarticulation and theories of intrinsic timing. *Journal of Phonetics*, 8, 113–133.

Garnham, A. (1982). Testing psychological theories about inference making. *Memory & Cognition*, 10, 341–349.

Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition*, 9, 434–443.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 311–325.

Gopher, D., & North, R. A. (1974). The measurement of attention capacity through concurrent task performance with individual difficulty levels and shifting priorities. In *Proceedings of the eighteenth annual meeting of the Human Factors Society*. Santa Monica, CA.

Gopher, D., & North, R. A. (1977). Manipulating the conditions of training in time-sharing performance. In *Proceedings of the 21st annual meeting of the Human Factors Society*. Santa Monica, CA.

Gould, A. & Stephenson, G. M. (1967). Some experiments relating to Bartlett's theory of remembering. *British Journal of Psychology*, 58, 39–49.

Graesser, A. C. (1981). Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, 13, 1–26.

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

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108, 356–388.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13, 512–521.

Hintzman, D. L., Nozawa, G., & Irmscher, M. (1982). Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior*, 21, 127–141.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110, 306–340.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.

Kinchla, R. A. (1980). The measurement of attention. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory*. New York: Academic Press.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Klapp, S. T. (1974). Syllable dependent pronunciation latencies in number naming, a replication. *Journal of Experimental Psychology*, 102, 1138–1140.

Kolodner, J. L. (November 1980). *Retrieval and organizational strategies in conceptual memory: A computer model* (Research Rep. No. 187). Yale University.

Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Psychology*, 7, 281–328.

Kolodner, J. L. (1984). *Retrieval and organizational strategies in conceptual memory: A computer model*. Hillsdale, NJ: Erlbaum.

Koriat, A., & Lieblich, I. (1977). A study of memory pointers. *Acta Psychologica*, 41, 151–164.

LaBerge, D. (1975). Acquisition of automatic processing in perceptual and associative learning. In P. M. A. Rabitts & S. Dornic (Eds.), *Attention and performance V*. New York: Academic Press.

Lachman, J. L., & Lachman, R. (1980). Age and the actualization of world knowledge. In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging: Proceedings of George Talland memorial conference*. Hillsdale, NJ: Erlbaum.

Lachman, R. (1973). Uncertainty-effects on time to access the internal lexicon. *Journal of Experimental Psychology*, 99, 199–208.

Lachman, R., Lachman, J., & Thronesberry. (1979). Metamemory through the adult life span. *Developmental Psychology*, 15, 543–551.

Lane, D. M. (1980). Incidental learning and the development of selective attention. *Psychological Review*, 87(3), 316–319.

Lehnert, W. (1977). Human and computational question-answering. *Cognitive Science*, 1, 47–73.

Lehnert, W. G. (1978). *The process of question answering*. Hillsdale, NJ: Erlbaum.

Lorch, R. F. (1981). Effects of relation strength and semantic overlap on retrieval and comparison processes during sentence verification. *Journal of Verbal Learning and Verbal Behavior*, 20, 593–610.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs*, 54, (No. 248).

Mandler, G. (1980). The judgment of previous occurrence. *Psychological Review*, 87, 252–271.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Pt. 1. An account of basic findings. *Psychological Review*, 88(5), 375–407.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. D. (1986). The appeal of parallel-distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: Bradford Books.

McLeod, P. (1977). A dual task response modality effect: Support for multiprocessor models of attention. *Quarterly Journal of Experimental Psychology*, 29(4), 651–667.

Moray, N. (1967). Where is capacity limited? A survey and a model. *Acta Psychologica*, 27, 84–92.

Moray, N., & Fitter, N. (1973). A theory and the measurement of attention: Tutorial review. In S. Kornblum (Ed.), *Attention and performance IV*. New York: Academic Press.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, 86(3), 214–255.

Navon, D., & Gopher, D. (1980). Interpretation of task difficulty in terms of resources: Efficiency, load, demand, and cost composition. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, NJ: Erlbaum.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgements for predicting perpetual identification and relearning. *Journal of Experimental Psychology*, 113(2), 301–323.

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(2), 279–287.

Nelson, T. O., & Narens, L. (1980). 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(6), 338–368.

Nickerson, R. S. (1977a). Crossword puzzles and lexical memory. In S. Dornic (Ed.), *Attention and performance VI*. New Jersey: Erlbaum.

Nickerson, R. S. (1977b October). Some comments on human archival memory as a very large data base. In *Proceedings on very large data bases*. Tokyo, Japan: Third International Conference on Very Large Data Bases.

Nickerson, R. S. (1980a). Motivated retrieval from archival memory. In J. H. Flower (Ed.), *Symposium on Motivation, 1980*. Lincoln: Univ. of Nebraska Press.

Nickerson, R. S. (1980b). Retrieval efficiency, knowledge assessment and age: Comments on some welcome findings. In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging*. Hillsdale, NJ: Erlbaum.

Norman, D. A. (1973). Memory, knowledge, and the answering of questions. In R. L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola Symposium*. Washington, DC: Winston.

Norman, D. A., & Bobrow, D. G. (1979). An intermediate stage in memory retrieval. *Cognitive Psychology*, 11, 107–123.

Norman, D. A., Rumelhart, D. E., & the LNR Research Group (1975). *Explorations in cognition*. San Francisco: Freeman.

Posner, M. I., Nissen, M. J., & Ogden, W. C. (1978). Attended and unattended processing modes: The role of set for spatial location. In H. L. Pick & I. J. Saltzman (Eds.), *Modes of perceiving and processing information*. Hillsdale, NJ: Erlbaum.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 93–134.

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review*, 83(6), 190–214.

Read, J. D., & Bruce, D. (1982). Longitudinal tracking of difficult memory retrievals. *Cognitive Psychology*, 14, 280–300.

Reder, L. M. (1976). *The role of elaborations in the processing of prose*. Doctoral dissertation, University of Michigan. Available through University Microfilms, Ann Arbor.

Reder, L. M. (1979). The role of elaborations in memory for prose. *Cognitive Psychology*, 11, 221–234.

Reder, L. M. (1982). Plausibility judgments vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, 89, 250–280.

Reder, L. M., & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. *Cognitive Psychology*, 12, 447–472.

Reder, L. M. & Dennler, C. (1984). The Moses illusion re-visited: factors affecting memory search. Unpublished data.

Reder, L. M. & Fabri, S. (1982). The role of topic knowledge on feelings of knowing: Time to estimate or answer and proportion attempted as a function of prior knowledge. Unpublished data.

Reder, L. M., & Ross, B. H. (1983). Integrated knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 9, 55–72.

Reder, L. M., & Wible, C. (1984). Strategy use in question-answering: Memory strength and task constraints on fan effects. *Memory & Cognition*, 12(4), 411–419.

Reiser, B. J. (1983). *Contexts and indices in autobiographical memory*. Doctoral dissertation, Yale University, Tech. Rep. 24, Cognitive Science Program.

Reiser, B. J., Black, J. B., & Abelson, R. P. (1985). Knowledge structures in the organization and retrieval of autobiographical memories. *Cognitive Psychology*, 17(1), 89–137.

Reiser, B., Black, J., & Kalamarides, P. (1986). Strategic memory search processes. In D. C. Rubin (Ed.), *Autobiographical memory*. New York: Cambridge Univ. Press.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12, 1–20.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4), 552–631.

Schank, R. C. (1975). *Conceptual information processing*. Amsterdam: North-Holland.

Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. New York: Cambridge Univ. Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing. I. Detection, search, and attention. *Psychological Review*, 84, 1–66.

Schweickert, R. (1978). A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology*, 18(2), 105–139.

Schweickert, R. (1980). Critical path scheduling of mental processes in a dual task. *Science*, 209, 704–706.

Schweickert, R. (1983). Latent network theory: Scheduling of processes in sentence verification and the Stroop effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(3), 353–383.

Shaw, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, NJ: Erlbaum.

Shiffrin, R. M. (1970). Memory search. In D. A. Norman (Ed.), *Models of human memory*. New York: Academic Press.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills: The eighteenth annual Carnegie Symposium on cognition*. Hillsdale, NJ: Erlbaum.

Singer, M., & Ferreira, F. (1983). Inferring consequences in study comprehension. *Journal of Verbal Learning and Verbal Behavior*, 22, 449–474.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 84, 214–241.

Snodgrass, J. G. & Townsend, J. T. (1980). Comparing parallel and serial models: Theory and implementation. *Journal of Experimental Psychology: Human perception and performance*, 6, 330–354.

Sperling, G., & Melchner, M. (1979). Visual search, visual attention and the attention operating characteristic. In J. Requin (Ed.), *Attention and performance VII*. New York: Academic Press.

Sternberg, S., & Monsell, S. (1981). *Speech programming: A critical review, a new experimental approach, and a model of the timing of rapid utterances*. Unpublished manuscript.

Townsend, J. T. (1971). A note on the identifiability of parallel and serial processes *Perception and Psychophysics*, 10, 161–163.

Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. In B. H. Kantowitz (Ed.), *Human Information Processing: Tutorials in performance and cognition*. Hillsdale, NJ: Erlbaum.

Townsend, J. T. (1976). Serial and within-stage parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, 14, 219–238.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press.

Whitten, W. B., & Leonard, J. M. (1981). Directed search through autobiographical memory. *Memory & Cognition*, 9, 566–579.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *and performance VIII*. Hillsdale, NJ: Erlbaum.

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very-long term memory. *Cognitive Science*, 5, 87–119.

Williams, M. D., & Santos-Williams, S. (1980). Method for exploring retrieval processes using verbal protocols. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, NJ: Erlbaum.

Zacks, R. T., Hasher, L., & Sanft, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 106–116.

(Accepted July 3, 1986)