Psychological Review
1982, Vol. 89, No. 3, 000–000

0033-295X/82/0000-0000$00.00

Plausibility Judgments Versus Fact Retrieval: Alternative Strategies for Sentence Verification

Lynne M. Reder
Carnegie-Mellon University

This article contrasts two views about how people judge the truth of statements. The more common view maintains that people decide whether a statement is true by finding a close (propositional) match to the query in memory; only if that fact cannot be found do they try to infer whether the statement is true by judging whether it is plausible. The second view, developed and argued in this article, is that judging plausibility is a more efficient strategy than direct retrieval (finding a propositional match), except when verbatim memory is very good. A model is proposed that exemplifies the second view. It is assumed that a person can evaluate a statement either by plausibility judgment or by direct retrieval. Both strategies consist of two major stages: searching for needed information and evaluating the adequacy of the retrieved information. Only when verbatim traces are strong, at very short delays after acquisition, is direct retrieval faster than judging plausibility. Direct retrieval becomes a less efficient strategy than plausibility judgment over time because the search stage becomes very long. Regardless of the ostensive task asked of a person, whether recognition or plausibility judgment, people use both strategies to answer questions. A person's preference for a particular strategy depends in part on task demands and in part on delay. Data are described from several experiments that support these theoretical positions and the data are fit by a formal model.

The processes involved in different types of question-answering tasks have been a topic of considerable interest. The types of tasks that have received the most attention are sentence recognition (e.g., Anderson & Paulson, 1977; Bransford & Franks, 1971; Dooling & Christiaansen, 1977; Hayes-Roth & Hayes-Roth, 1977; Kintsch & Bates, 1977; Sachs, 1967) and sentence verification (e.g., Carpenter & Just, 1975; Clark & Chase, 1972; Kintsch, 1974; Trabasso, Rollins & Shaughnessy, 1971). Experiments that ask subjects to judge whether they recognize a test probe typically require a match of the probe with a structure in memory. The process underlying this kind of question-answering task will be called *direct retrieval.* The sentence verification task, on the other hand, typically involves at least some minimal computation or inferring process.

There has been a fair amount of speculation concerning the relative efficiency of direct matching as compared with inferential reasoning (e.g., Camp, Lachman, & Lachman, 1980; Collins & Loftus, 1975; Collins & Quillian, 1969; Haviland & Clark, 1974; Kintsch, 1974; Lachman, 1973; Lachman & Lachman, 1980; Lehnert, 1977). Virtually all viewpoints are in agreement on the assertion that a person's preferred strategy for question answering is direct retrieval. Others, through their simulations, have implied the same thing (e.g., Anderson, 1976; Anderson & Bower, 1973; Norman, Rumelhart, & the LNR Research Group, 1975; Quillian, 1968; Schank & Abelson, 1977). Lachman and Lachman (1980) articulate this commonly held conception of the relationship between fact retrieval and the drawing of inferences:

> 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

The author wishes to acknowledge the helpful comments of R. Jeffries, J. Larkin, M. Schustack, and H. Simon on an earlier version of the manuscript and is especially grateful to J. R. Anderson for much useful advice in all stages of the research.

Requests for reprints should be sent to Lynne M. Reder, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

> 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)

Despite the wide speculation on the advantage of direct retrieval, there have been only a few experiments that contrast a subject's ability to judge truth or plausibility with recognition ability (e.g., Kintsch, 1974; Singer, 1979b), and these have not explicitly compared subjects' speed at making a recognition judgment with speed at making a plausibility judgment. This article will argue that plausibility judgments are faster in delayed tests when the relevant information is not highly available. Therefore, it will be useful to investigate how plausibility judgments compare with recognition judgments at various delay intervals.[1]

Kintsch (1974) argued that we can verify statements that are inferable from information read as fast as statements that were explicitly asserted if the inferences are made and stored at the time of study. In these experiments (performed by McKoon and Keenan; Kintsch, 1974, chap. 8) subjects read short passages where statements were either directly asserted (called *explicit statements*) or clearly followed from the text and were necessary for comprehension (*implicit statements*). After subjects finished reading the passage, they were asked to verify explicit and implicit statements immediately or were asked to verify these statements 15 minutes later. In the immediate condition subjects were faster at verifying explicit than implied statements. However, with a 15-minute delay, the difference in response time between the explicit and implicit conditions was no longer significant, although there remained a slight advantage for the explicit condition. From these data Kintsch argued that implied statements are inferred during comprehension and that inferential statements are verified by fact retrieval, that is, by searching for the specific proposition in memory. Once the lexical-trace advantage of the explicit statements is gone (15 minutes after reading), there should be no difference because in both cases the exact proposition is stored in memory and is recovered for question answering. Note that these data and this view are not really inconsistent with the view of Lachman and Lachman (1980). In this case inferences are also verified by direct retrieval.

Kintsch (1974, chap. 9, with Monk) compared true/false judgments (the true sentences were stated in the text) with recognition judgments. Subjects were tested immediately after reading the material and true/false judgments were found to be slower than recognition judgments. Kintsch concludes that both types of judgments involve basically similar processes, namely, checking whether a certain memory trace is or is not available in memory. The reason true/false judgments are slower, he argues, is that they require accessing propositional memory whereas recognition judgments can rely on surface features.

Haviland and Clark (1974) and Clark and Haviland (1977) present data showing that subjects take longer to comprehend a statement if a referent in the sentence has to be inferred from preceding information. In their experiment, subjects read sentence pairs. One sentence is presented at a time and the subject's task is to push a button when the presented sentence is understood. The dependent measure is time to comprehend the second sentence of the pair. This sentence is constant across conditions; only the preceding sentence varies. For example, "Fran took the picnic supplies out of the car. The beer was warm," involves an inference or "bridge." To comprehend the second sentence with respect to the first, the subject must infer that one of the picnic supplies was beer. In contrast, the pair "Fran took the beer out of the car. The beer was warm." involves no inference. Time to comprehend the second sentence is faster in this case. (Haviland and Clark controlled for the double presentation of "beer.") The Kintsch (1974) view is not inconsistent with the result that the drawing of an inference takes

[1] Kintsch and his colleagues (e.g., McKoon & Keenan cited in Kintsch, 1974) have looked at a verification task at several delay intervals, but they did not compare verification times with fact recognition times. Monk and Kintsch (in Kintsch, 1974) contrasted true/false judgments with recognition judgments but did not vary delay.

longer than fact retrieval when the task is to determine reference. Such a task necessarily occurs moments after the information is presented. In the immediate condition of the Kintsch experiments, too, verification was faster for explicit statements than for implicit statements.

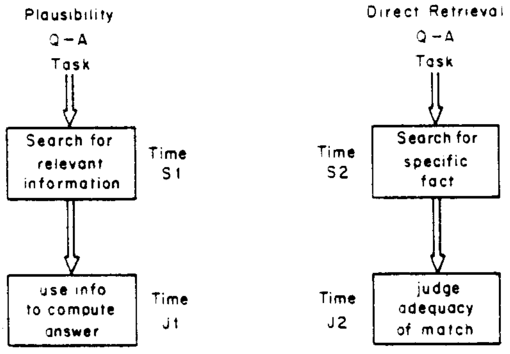
Singer (1979a, 1979b), like Haviland and Clark (1974), has evidence that inferences are slower than direct retrieval questions. In a sentence verification task that occurred either immediately after reading or 20 minutes later, he found that subjects were slower to respond affirmatively to not-presented, but implied, sentences than to ones that quoted or paraphrased the passage. From this he concluded that at least some of the cognitive processes associated with making the inferences to answer a question are executed at test time. He did not conclude, as Kintsch (1974) had, that responses to inferences are initially slower than to explicit statements due to the lack of a lexical trace. He ruled this out because his subjects responded faster to paraphrases than inferences. However, Kintsch's implied statements were required for textual coherence whereas Singer's were not. This may account for why the statements Kintsch tested were inferred during reading.

Reder (1979) also presented evidence consistent with the view that subjects compute inferences at time of test even when the information is stored in memory. In those studies the plausibility of the test sentence (with respect to the story being queried) affected plausibility judgment time even when the item had been explicitly presented. When the test sentence had been "primed" earlier by asking the subject to answer a related question while reading the story, there was also a plausibility effect. The decision times were faster for test items that had been previously presented or primed than for those not treated, suggesting that the manipulations had an effect. However, because there was a large effect of statement plausibility for explicit and primed statements, it does not seem that direct retrieval is always tried first. Reder (1979) argued that subjects were faster in the explicit condition because there was more information from which to make a plausibility judgment. Thus, the fact that explicitly presented statements are verified faster does not have to mean that the process of direct retrieval is faster than that of inference.

The position that will be argued in this paper is somewhat heretical: I believe that fact retrieval (trying to find an assertion in memory) is often less efficient than computing plausibility (or inferring) and that it is not always the first strategy employed in sentence verification. This view is not based solely on the Reder (1979) data. Rather, the position seems reasonable on a number of counts. In everyday life it is unlikely that all facts or even the majority of facts on which people are queried are directly stored in memory. Further, memory is a rich, highly redundant store of information. Searching for any specific proposition may not be much easier than searching for a needle in a haystack. Therefore, it is often faster to select the first few relevant facts found in memory (and compute the answer) than to continue to search until an exact match can be found. In some cases it is fairly intuitive that computing plausibility should be easier. For instance, if asked to judge whether the boys in William Golding's *Lord of the Flies* were savage, we do not try to retrieve the exact proposition that the boys were savage. Rather we sample from the rich set of facts we know about the novel that would seem relevant and then judge whether the probe seems plausible.

One could argue that the only reason plausibility judgments might be faster than recognition judgments is that there are two ways to decide that an explicit statement is plausible, namely, plausibility or direct retrieval. That is, if one were to assume a parallel-race model where the two question-answering strategies are tried in parallel, then when whichever process is completed first a decision could be made. The position argued for here, however, does not rest on the assumption of a parallel race. In fact, I want to argue that judging plausibility can be a faster question-answering process even when the test statement has not been previously presented. If it can be shown that subjects are faster to judge a statement as plausible when it was not stated than to recognize a sentence that was stated, then the result cannot be explained merely by assuming that there are two ways to judge plausibility but only one way to recognize.

Below more formal arguments are developed for reaction time and accuracy predictions, comparing direct retrieval with judging plausibility. These predictions can be tested in a number of ways. Several experiments are reported, and the resulting pattern of data is compared to theoretical predictions. Then quantative fits are made to the data using the described model and compared with fits to the data using the commonly assumed model of direct retrieval first. The parameter estimates derived from the fit of the proposed model are evaluated using several criteria. Other studies, published and unpublished, are also shown to provide further empirical support.



J1 > J2 S1 ≤ S2

When delay is short: S1 + J1 > S2 + J2

When delay is long: S1 + J1 < S2 + J2

Figure 1. Schematic model of stages involved in fact retrieval and plausibility judgment types of question answering. (Relative time parameters derive from duration of the various stages.)

Plausibility Judgments Versus Direct Retrieval

Statements can be verified by one of two processes: either by direct match with the same assertion in memory or by computing plausibility. Figure 1 illustrates schematic models of these two types of judgment processes. For both types it is assumed that first a person must find the appropriate information in memory and then evaluate it. This means that each process contains two stages, denoted as the search stage and decision stage. The time to complete the search stage is *S1* for the plausibility task and *S2* for the recognition task. The time to compute plausibility (in the decision or judgment stage) has mean *J1*, and the time to evaluate the adequacy of the retrieved fact in the recognition task has mean *J2*.

Clearly *J1* will be greater than *J2* because the decision process involved in the plausibility task is more complicated than matching the retrieved fact with the test probe.[2] I suspect that it is because of this difference in the judgment stages of the two processes that most theorists argue that inferential question answering takes longer. The reason plausibility judgments can be more efficient is because the search stage involved in both tasks is often faster for the inferential or plausibility task. Direct retrieval relies on finding in memory a specific fact that may be relatively unavailable or inaccessible. Plausibility judgments accept any number of possible facts to use for computing plausibility. There are many relevant facts in memory that can be used for computing plausibility because people have elaborated or embellished the input when they comprehended it.

The assumptions about the retrieval stage can be interpreted in terms of a semantic network representation of a person's factual knowledge (much like that of Anderson, 1976; Collins & Loftus, 1975; or Norman, Rumelhart, & the LNR Research Group, 1975) and a spreading-activation process operating on this network (e.g., Anderson, 1976; Collins & Loftus, 1975). A semantic network consists of interconnected propositions that are joined at concept nodes. Activation spreads out of nodes (that were activated by a test probe) and travels down the relational arcs connecting them to other nodes and propositions. How much activation goes down a given link (or how fast activation spreads) is a function of that link's

[2] Reder (1976) describes potential plausibility judgment mechanisms. The exact nature of these mechanisms is not critical to the current discussion.

strength. A link's strength is in part a function of its *recency* of past activation (the last time a person thought about it) and *frequency* of activation (how often the person has thought about it) and in part a function of the strength of the other propositions connected to the same node that must share the activation.

The reason direct retrieval slows with delay is that a specific fact loses the strength benefit of recency and so will take longer to be activated. The search process used in judging plausibility, on the other hand, does not rely on any specific fact. Assuming that the mean time to find any particular fact in memory has a probabilistic distribution and that the first acceptable facts stop the search process (a race for a subset to be activated), search will stop sooner for "plausibility search."

The poorer memory is for the topic under query, the greater is the advantage of the

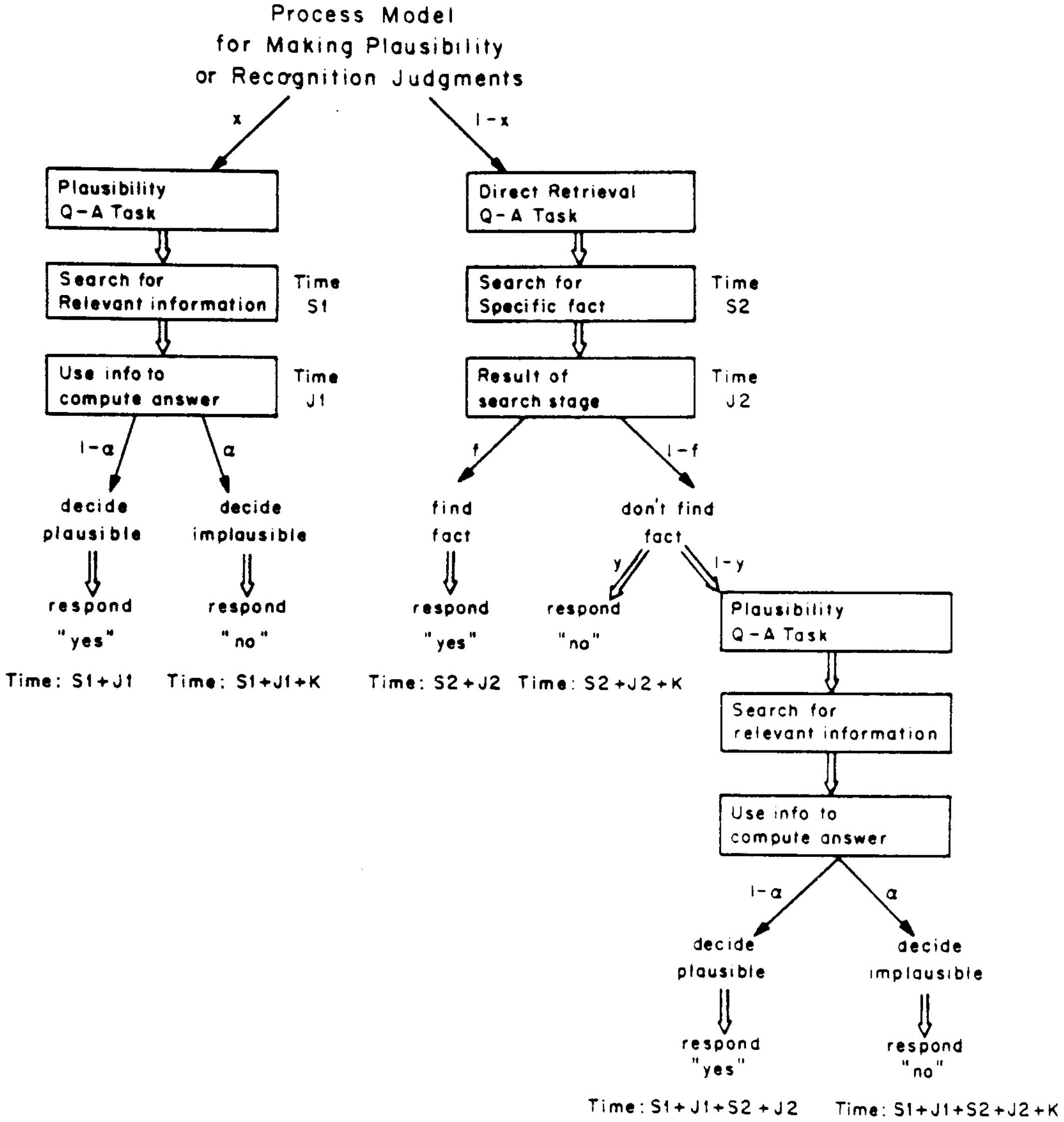


Figure 2. Probabilistic model of alternative strategies employed to judge the plausibility of or recognize a statement.

plausibility search stage over the fact retrieval search stage. If all facts are highly accessible, there is no advantage for the looser search criterion of the plausibility judgment task. That is, search time for plausibility, *S1*, is much less than *S2* when memory traces are weak but is essentially the same when memory traces are very strong (a ceiling effect). Given that the decision stage for plausibility is slower than the one for recognition, and does not depend on delay, time to judge plausibility will be longer than recognition decisions at short test delays. However, at longer delays, when information is less accessible, the advantage of *S1* over *S2* will more than compensate for the other inequality and plausibility judgments will be faster.

The above line of argumentation is an intuitive explanation of why one type of process can be more or less efficient than the other. However, I do not believe that all people always perform the task that is asked of them. Rather, a person's strategy is a function of both the task demands and the situational context. The fact that direct retrieval is initially more efficient than. plausibility may affect preference and, likewise, if it is true that ultimately plausibility is an easier mode for question-answering than is direct retrieval, this too may affect strategies, despite the ostensive task requirements.

The simple model of Figure 1 has been complicated in Figure 2. This 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) allows one to predict reaction times and error rates to answer questions as a function of task, delay, and plausibility.

The assumption that a person does not always answer a question by retrieving a direct match or by computing plausibility, regardless of the ostensive task, is reflected in the first pair of branches. With probability x, a subject answers the question by deciding whether it seems plausible. The value of x varies both as a function of the task the person was actually asked to perform and the delay between acquiring the information and the testing. When the test is right after learning the material, the memory traces are quite strong and a person has confidence in his or her ability to find the statement in memory, if it has been presented. Therefore, it is assumed that x increases with delay, because memories will fade. It is also assumed that x is greater for people asked to judge plausibility than for those asked to recognize. When the plausibility branch, or strategy, is selected, the time to search for relevant information, *S1*, will not vary with the "official" task. *S1* is also assumed to remain relatively constant across delay and plausibility of the probe, yielding the long-term advantage for plausibility judgments over direct retrieval.

Once relevant information is found, the probability, α, of deciding that a statement is not plausible depends on the statement's inherent plausibility. Highly plausible statements have low values of α. The time to compute whether a statement is plausible, *J1*, also varies with the plausibility of the statement, taking longer for less plausible statements that are still judged to be plausible.

On the other hand, when a person tries to retrieve a specific fact $(1 - x)$, search will be successful with probability f, if in fact the probe had been previously stated. The value of f varies with delay between learning the statement and being asked to judge it. If the probe is not found in memory, then with probability y the person gives up and responds negatively. Alternatively, with probability $1 - y$, he or she will try to answer the question by judging plausibility. The probability is very high of going on to use plausibility mechanisms when direct retrieval fails if the person was actually asked to evaluate the plausibility of the statement, that is, y should also vary with task. When the task requires a recognition judgment, the likelihood of reverting to plausibility following a retrieval failure may depend on the person's confidence in the relevant memory. Of course, the probability of trying direct retrieval first before trying plausibility should also be affected by the strength of the memory traces. The retrieval parameter, *S2*, should vary with delay as the expected time to activate the exact fact increases. The search stage for recognition judgments has a much more stringent criterion for accepting a candidate to evaluate than does the plausibility search stage, which can take the first set of related, activated propositions. On the other hand, the evaluation time of a statement, *J2*, is fairly short in recognition.

A search for a close match to the probe will either exceed a cutoff time or fail to match on closer inspection with probability $1 - f$. When the person elects to try the plausibility strategy, with probability $1 - y$, the remaining processes are exactly the same as they would have been if plausibility were tried first. Note that the entire $1 - x$ branch represents the traditional view of how verification or plausibility judgments are made. The difference between the current view and traditional views is that some of the time people are assumed to try inferential or plausibility processes first.

When a statement's plausibility is being evaluated, the same probabilities (values of α) apply regardless of whether direct retrieval was tried first. Similarly, there is no saving in performing the plausibility tasks, $S1 + J1$, by trying direct retrieval first.

The reaction time (RT) predictions for correct responses can be computed as follows: Each path through the question-answering tree has a processing time associated with its terminus, which is the sum of the processing times of the stages on its path. Each path has a probability equal to the product of the probabilities along that path. The overall reaction time is just the weighted average of the individual path times, where the weights are the path probabilities.

The expected (E) time to say "yes" to a previously presented statement, then, is the weighted average of ecĥ time for saying "yes," multiplied by its respective probability, divided by the probability of saying yes.

$$\mathrm{E}(\mathrm{RT}/\text{"yes"}) = \frac{x(1-\alpha)(S1+J1) + (1-x)[f(S2+J2) + (1-f)(1-y)(1-\alpha)(S1+J1+S2+J2)]}{x(1-\alpha) + (1-x)[f + (1-f)(1-y)(1-\alpha)]}. \quad (1)$$

The value of these parameters in Equation 1, of course, would depend on various factors; for example, f depends on delay between acquisition and query, as would x. The value of x and y would also depend on the requested judgment type. The expected time to say "no" for a not-presented statement in a recognition task is computed in an analogous fashion. A constant, K, is added for negative responses. It is the sum of each time for saying "no," multipled by its respective probability, divided by the probability of saying "no."

$$\mathrm{E}(\mathrm{RT}/\text{"no"}) = \frac{x\alpha(S1+J1+\mathrm{K}) + (1-x)[y(S2+J2+\mathrm{K}) + (1-y)\alpha(S1+J1+S2+J2+\mathrm{K})]}{x\alpha + (1-x)[y + (1-y)\alpha]}. \quad (2)$$

The expected time to judge that a not-presented statement is plausible can be expressed as

$$\mathrm{E}(\mathrm{RT}/\text{"yes"}) = \frac{x(1-\alpha)(S1+J1) + (1-x)(1-y)(1-\alpha)(S1+J1+S2+J2)}{x(1-\alpha) + (1-x)(1-y)(1-\alpha)}. \quad (3)$$

assuming that $f = 0$, that is, people never find facts that were not presented. If, however, we assume that people can erroneously believe that they retrieved a fact that was not presented, we must add the parameter γ, the probability of finding a not-presented statement. This parameter replaces f in Figure 2 for situations where the probe had not been previously stated. Therefore Equation 2 is rewritten as Equation 4, and Equation 3 is rewritten as Equation 5 to reflect the additional parameter:

$$E(RT/\text{"no"}) = \frac{x\alpha(S1 + J1 + K) + (1 - x)(1 - \gamma)[y(S2 + J2 + K) + (1 - y)\alpha(S1 + J1 + S2 + J2 + K)]}{x\alpha + (1 - x)(1 - \gamma)[y + (1 - y)\alpha]} \quad (4)$$

$$E(RT/\text{"yes"}) = \frac{x(1 - \alpha)(S1 + J1) + (1 - x)[\gamma(S2 + J2) + (1 - \gamma)(1 - y)(1 - \alpha)(S1 + J1 + S2 + J2)]}{x(1 - \alpha) + (1 - x)[\gamma + (1 - \gamma)(1 - y)(1 - \alpha)]}. \quad (5)$$

The model also makes predictions about error rates. The probability of an error is the sum of the probability values for each path that leads to a wrong response. The probability value of a path is the product of each probability associated with a branch of that path. Consider a recognition task. There are several ways for a subject to make an error to a presented statement. With probability x the subject chooses to judge plausibility and with probability α the statement is erroneously judged as implausible. With probability $1 - x$ the subject chooses to try direct retrieval but with probability $1 - f$, the subject does not find the presented statement. In this case the subject may stop and say "no" (probability y) or go on and judge the statement as implausible (probability $[1 - y]\alpha$). Therefore, the probability (P) of an error (e) of saying "no" in a recognition task to a presented statement is

$$P(e) = x\alpha + (1 - x)[(1 - f) \times (y + (1 - y)\alpha)]. \quad (6)$$

The probability of an error by saying "yes" to a not-presented statement in the recognition task is similar (note that γ replaces f):

$$P(e) = x(1 - \alpha) + (1 - x) \times [\gamma + (1 - \gamma)(1 - y)(1 - \alpha)]. \quad (7)$$

In a plausibility judgment task, the way to make an error to a previously presented statement is to decide erroneously that a plausible statement is implausible or to fail to find it and not go on to judge plausibility. Therefore, the probability of error is

$$P(e) = x\alpha + (1 - x)(1 - f) \times [y + (1 - y)\alpha]. \quad (8)$$

For nonexplicit inferences, f is replaced by γ:

$$P(e) = x\alpha + (1 - x) \times (1 - \gamma)[y + (1 - y)\alpha]. \quad (9)$$

Testable Hypotheses

The theoretical framework presented above leads to a number of expectations about performance that can be tested empirically. Each of these expectations, if confirmed, has an important theoretical implication.

1. When information about a story is not highly available, a statement that is true with respect to a story would be judged plausible faster than it would be recognized. If it is true that plausibility is faster in these situations, then people must not always try direct retrieval first.

2. As memory traces weaken, plausibility judgments become faster than recognition judgments. This suggests that there are two strategies with shifting propensities to be employed first.

3. People are faster to judge a statement as plausible even when it had not been presented in a text than to recognize that statement when it had been presented. This argues against explaining the plausibility task advantage by assuming a race between the two strategies operating in parallel. Prediction 2 also argues against a parallel-race model.

The two experiments to be reported tested these qualitative predictions about performance. The model just sketched will be fitted to the specifics of the experiments.

Experiment 1

The task in this experiment was quite simple. Subjects read short, mildly interesting stories and then were asked questions about them. Some subjects were asked to make judgments concerning whether a test probe had been presented in the story. Other subjects were asked to judge whether the test probe was plausible, given the story. This variable was manipulated between subjects.

For subjects asked to make plausibility judgments, half of the test items (probes) were implausible, included only to keep the probability of a positive response at 50%. Both groups of subjects were tested on an equal number of highly plausible and moderately plausible statements. For both groups of subjects, half of the moderately and highly plausible statements were explicitly stated in the stories. (Implausible statements were never stated in the story.) Implausible statements were not included as test items in the recognition condition, because this would have made subjects more inclined to adopt a strategy of judging plausibility rather than doing the prescribed task.

The fourth factor in the experiment was the delay between the presentation of the material to be tested and the test itself. Questions could be asked after each story, after reading all (10) stories, or 2 days later. This was also a between-subjects factor.

Method

Materials. Ten stories written by five different authors were used. The questions about the stories and the stories themselves had been previously used (Reder, 1976, 1979). Examples of the materials and more detail about material construction can be found in Reder (1976, 1979).

Questions to be judged affirmatively in the plausibility task varied on the dimension of plausibility: half were highly plausible and half moderately plausible, but still clearly plausible relative to the foils. Plausibility of a statement was defined by other subjects' ratings of the inferences after reading the story. The statements to be rated had not been stated in the story. Presumably, this fact makes little difference for highly plausible statements. However, moderately plausible statements are assumed to become more plausible when stated in the story. The implausible statements used in the plausibility judgment condition contained the same concepts used in the story so they could not be rejected on the basis of lexical familiarity. The statements were also not implausible without reading the story.

Procedure and design. Subjects read 10 stores at their own pace. They were told to read the stories in a normal fashion, as they would when reading for pleasure, and that later they would be asked some questions about the stories. One sentence of a story at a time was presented on the computer controlled video screen.

Subjects were randomly assigned to either the recognition judgment task or the plausibility judgment task. Until the subject was asked the first set of questions, there was no difference in the procedure or materials for the two groups. The plausibility judgment group was asked to decide if a statement seemed true given the story they read, whereas the sentence recognition group was asked to judge if a particular statement had actually been presented in the story.

The test probes for the recognition task varied on two orthogonal dimensions: stated versus not stated in the story and plausibility (highly or moderately plausible). Only the first dimension defined how subjects in the recognition task should respond to a probe. Subjects assigned to the plausibility task saw these same four types of probes; however, all would be responded to positively. Plausibility subjects were also tested on an equal number of implausible statements so they would have an equal proportion of positive and negative test items.

Different groups of subjects were asked questions about the stories at one of three delay intervals: after each story, after all 10 stories (approximately 20 minutes later), or 2 days after reading the stories. The first line of each story was preceded by its title so that when questions about a story were asked at a delay, the subject knew which story was being queried by first seeing its title.

Subjects. Twenty-seven subjects were used in the immediate delay condition: 14 in the recognition task and 13 in the plausibility task. Thirty-two subjects participated at the 20-minute delay: 15 in recognition and 17 in plausibility. They received one credit towards a course requirement. There were 60 subjects in the 2-day delay, 30 in each task type. Because they had to return for a second session, these subjects received two credits, one credit and $2.50, or $5.00.[3]

Results

Table 1 displays the mean response times in seconds for correct responses and error rates for both the plausibility task and the recognition task at each level of delay in Experiment 1. The data are broken down according to those probes that had been stated or not stated in the story and whether they are highly plausible or moderately plausible. Performance on the implausible statements is also given for subjects in the plausibility task. Figure 3 displays the response times for correct decisions for plausible

[3] The three levels of delay in this experiment were actually separate experiments. Subjects were randomly assigned to the two judgment tasks, but it was impossible to randomly assign subjects to delay because one level required subjects to return 2 days later. Rather than run all conditions concurrently, only one experiment was available at a time so that subjects would not self-select into or out of the 2-day, better paying experiment. In Experiment 2 the three levels of delay were run in a different order within a semester.

statements as a function of task delay and whether the probe had been stated in the story prior to test. The three levels of delay—immediate (right after reading the story, which was approximately 2.5 minutes), 20 minutes (after reading all 10 stories), or 48 hours later—are indicated logarithmically on the abscissa. The ordinate represents response time in seconds. In this graph the data are collapsed over the plausibility of the statements.

A 2 × 2 × 2 × 3 analysis of variance was performed on the correct response times and on the percentage of correct responses for the factors of plausibility, presentation (stated and not-stated), and judgment task (recognition vs. plausibility) and three levels of delay (immediate, 20 minutes, and 2 days). The first factors are within-subjects whereas the latter two factors are between-subjects variables. The error term used was always the interaction term of the effect of interest with subjects.

Of the 30 possible main effects and interactions, 20 were significant. Some of the significant effects and interactions, both for error rates and response times, are noteworthy: subjects were less accurate with delay, $F(2, 112) = 35.64$, $p < .001$; accuracy was much worse for recognition subjects, $F(1, 112) = 98.56$, $p < .001$; and the difference in accuracy between recognition and plausibility subjects increased with delay, $F(2, 112) = 14.57$, $p < .001$. There was no effect of delay on response time because plausibility judgments were getting faster and recognition judgments were getting slower. This interaction of task with delay on RT was significant, $F(2, 112) = 4.1$, $p < .02$. The plausibility of the probes had an effect on latency, $F(1, 112) = 40.92$, $p < .01$, such that subjects respond faster to highly plausible inferences. They also respond faster, $F(1, 112) = 113.22$, $p < .01$, and more accurately, $F(1, 112) = 122.06$, $p < .01$, to statements that were stated in the text. The triple interaction of task, plausibility, and whether the probe had been stated on speed, $F(1, 112) = 18.71$, $p < .01$, and on accuracy, $F(1, 112) = 55.15$, $p < .01$, reflects the fact that there is a plausibility effect in the recognition task.

A number of results are highlighted in Figure 3. First, reaction times in the recognition task increase with delay. It is not surprising that subjects should take longer to make these judgments because memory presumably becomes poorer. A less intuitive result, however, is the finding that response times become shorter with longer delays in the plausibility judgment task, the greatest speedup being for not-stated items from the

Table 1
Mean Response Times (in seconds) and Error Rates for Experiment 1

	Judgment task			
	Recognition		Plausibility	
Delay	Stated	Not stated	Stated	Not stated
Immediate				
High plausibility	2.28 (.18)	2.70 (.21)	2.66 (.03)	3.29 (.08)
Medium plausibility	2.38 (.14)	2.68 (.14)	2.82 (.08)	4.04 (.23)
Implausible			3.51 (.07)	
20 minutes				
High plausibility	2.48 (.13)	2.67 (.57)	2.52 (.09)	2.54 (.14)
Medium plausibility	2.66 (.19)	2.77 (.24)	2.58 (.13)	3.08 (.29)
Implausible			2.79 (.13)	
2 days				
High plausibility	2.52 (.16)	3.12 (.68)	2.41 (.09)	2.52 (.13)
Medium plausibility	2.80 (.21)	3.06 (.51)	2.60 (.16)	2.89 (.25)
Implausible			2.62 (.17)	

Note. Error rates are in parentheses.

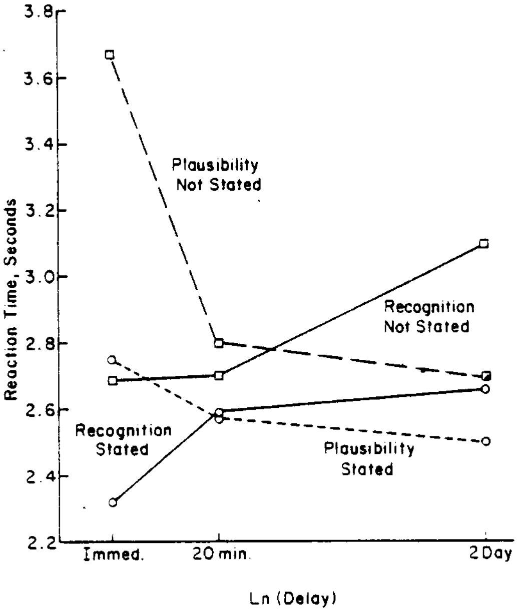


Figure 3. Mean reaction time in Experiment 1 for plausibility and recognition judgments as a function of whether the probe had been stated in the story, plotted across levels of delay.

immediate to 20-minute delay. Subjects are initially slower at making plausibility judgments than at making recognition judgments but are faster 2 days after reading the stories. Initially subjects are only slightly more accurate (7%) in the plausibility judgment task, but the accuracy of plausibility judgments is hurt much less by delay than are the recognition judgments. At a 2-day delay, there is a 23% advantage for plausibility subjects (regardless of whether accuracy on implausible statements is included) and recognition performance is close to chance.

Discussion

The large speedup from the immediate condition to the 20-minute delay in the plausibility task for not-presented statements has an explanation related to that for the slowdown in the recognition task. The reason subjects are initially very fast in the recognition conditions may be due to the availability of the verbatim traces of the stories. The surface-structure information has faded by the 20-minute delay condition and this can account for why recognition performance becomes slower and less accurate. It is more difficult to search for a propositional match than to match strings of words. The seductive quality of matching verbatim traces may account for the very slow performance immediately in the not-presented (not-stated) plausibility condition. That is, even in the plausibility tasks, subjects might have tried to answer questions by matching word for word. If subjects did this in the not-presented plausibility condition, they would fail and would have to then try to answer the question using a plausibility judgment mechanism.

This explanation is essentially an embellishment of part of the model given earlier. Initially subjects are more inclined to try direct retrieval in the plausibility task as well as in the recognition task due to the strong verbatim traces. This strategy hurts the not-presented statements in the plausibility task because they perform a fruitless search prior to the required task. In the longer delay conditions when subjects are less inclined to try direct retrieval first, the not-stated condition of plausibility is greatly facilitated.

It was mentioned in the introduction that one reason plausibility judgments for presented statements might be faster than recognition judgments is that subjects had two ways to respond "yes," assuming a race between the two types of judgments. This parallel-race model seems less viable as an explanation of the plausibility advantage because initially subjects are slower in the plausibility task.

One has to be careful in the interpretation of any advantage of plausibility judgments over recognition judgments because different foils were used in the two tasks. The foils used in the recognition task were all plausible sentences not studied, whereas the foils used in the plausibility task were implausible sentences not studied.

It is clear that the judgment times in the plausibility task will vary with foil difficulty. If the items used as foils in the plausibility task were a string of *w*s or nonsense words, clearly the task would be trivial and would not involve judging plausibility. One could also construct foils that would be very difficult for a subject to discriminate by making foils more and more plausible with respect to the story (causing more errors). In this case plausibility judgments would always be slower than recognition judgments. Given that the foils were not comparable, is there anything that can be concluded from the relative times of recognition and plausibility judgments?

Despite this problem of foil comparability, there is something significant to be concluded from the fact that plausibility judgments are faster. The implausible foils used in this experiment were not a string of *ws* and they could not be rejected simply by understanding the nature of the task. Rather, a subject would have to read and understand a foil and evaluate it with respect to the previously read passage in order to decide that it was implausible. The fact that foils could only be rejected using a plausibilitylike mechanism and that in some conditions this strategy was faster than recognition implies that judging plausibility can be a faster strategy than making recognition decisions. This conclusion, although not as strong as one would like, does serve to rule out some models of fact verification.

The error-rate data tell an equally interesting and consistent story: Accuracy drops off substantially in the recognition conditions with the passage of time. Most of the drop in accuracy reflects an increase in acceptance of moderately and highly plausible foils. This is especially true for highly plausible statements. Subjects respond with false alarms to 70% of the highly plausible foils at the 2-day delay. Even for moderately plausible foils, recognition subjects are performing at 50% accuracy. This high rate of false acceptances of foils can mean that many of the correct acceptances of targets are basically guesses and, hence, are faster than they should be. The notion that higher error rates occur with faster reaction times has been called the "speed-accuracy tradeoff" (see Pachella, 1974, for a complete discussion.) This view holds that a person can alter his or her speed of responding by setting a different accuracy criterion or vice versa.

In contrast to the huge drop in accuracy in the recognition task, accuracy for plausibility judgments is affected much less by delay. This means that the mechanisms used to make plausibility judgments are almost as good 2 days later as they are immediately. The results of the recognition and plausibility conditions together indicate that, with time, we can no longer remember exactly what we were told, but our ability to judge the truth of an assertion, given what we were told, remains largely intact. Therefore, it seems that the more accurate method of judging statements at a delay is by computing plausibility not by searching for the assertion in memory.

Signal-detection analyses. The logic of the preceding argument can be supported by a formal analysis using the theory of signal detection (Swets, Tanner, & Birdsall, 1961) to separate true discrimination from response bias. This analysis requires two distributions: a signal distribution and a noise distribution. In the case of recognition, the signal distribution corresponds to the presented statements and the noise distribution to the not-presented statements. The value of d' corresponds to the distance between the means of these two distributions, measured in z scores or standard deviations of the standard normal. The greater the distance, the easier it is to discriminate targets from foils. Using similar assumptions one can also obtain an estimate of response bias, β, which is the likelihood ratio of the target distribution to the foil distribution at the criterion. (see Healy & Kubovy, 1978, for a fuller discussion about the usefulness of these measures.) There were actually two signal distributions—one for highly plausible presented and one for moderately plausible presented—and two noise distributions for corresponding not-presented statements.

For plausibility judgments the signal distribution corresponds to the plausible statements and the noise corresponds to the implausible statements. In this case there are four different signal distributions (two levels of plausibility × two levels of presentation) that are all measured against the same noise distribution.

Table 2 gives the values of d' and β as a function of delay, plausibility task, listing values separately for presented statements in the plausibility task, and for not previously presented statements in that task. Analyses of variance were done on d' and β. The results are quite straightforward and consistent with the discussion given above.

Table 2
d′ and β Values for Experiment 1

	Judgment task					
	Recognition		Plausibility stated		Plausibility not stated	
Delay	*d′*	*β*	*d′*	*β*	*d′*	*β*
Immediate						
Medium plausibility	2.35	1.38	3.10	1.42	2.41	3.01
High plausibility	2.01	2.20	3.46	.77	3.07	1.42
20 minutes						
Medium plausibility	1.71	1.40	2.32	1.24	1.75	1.82
High plausibility	.89	.66	2.63	.88	2.31	1.28
2 days						
Medium plausibility	.88	.77	2.14	1.16	1.75	1.36
High plausibility	.58	.70	2.44	.70	2.23	.84

For the recognition task *d′* is always bigger (i.e., discriminability is better) for moderately plausible than highly plausible statements, $F(1, 55) = 21.86$, $p < .01$. This is probably because of the tendency to use plausibility judgments even in the recognition task. (The triple interaction of task, plausibility, and whether the probe was stated reported in the Results section is consistent with this view.) Delay also had a large effect on *d′* in the recognition task, $F(2, 55) = 37.68$, $p < .01$, but plausibility and delay did not interact.

In contrast to the recognition results, when judging plausibility is the specified task, *d′* is always larger for highly plausible statements, $F(1, 58) = 68.87$, $p < .01$. The value of *d′* was greater in the stated than in the not-stated condition, $F(1, 58) = 44.3$, $p < .01$, and plausibility interacted with whether it was stated $F(1, 58) = 5.1$, $p < .05$. Delay also affects *d′* for plausibility, $F(2, 58) = 17.7$, $p < .01$. However, the value of *d′* reaches asymptote for plausibility at the 20-minute delay but continues to drop for recognition.

The measure of response bias, *β*, also provides converging evidence for the data described earlier: Overall, there was no effect of plausibility on *β* in the recognition task, but *β* did go from a value greater than 1.0 to less than 1.0 with delay, $F(2, 55) = 4.6$, $p < .05$, suggesting that there was a shift in tendency to accept more foils as well as a drop in discriminability. This tendency to lower the criterion to accept foils at longer delays was much greater for the highly plausible statements, $F(2, 55) = 5.5$, $p < .01$.

In the plausibility task the foils did not change as a function of plausibility or whether an item was stated; therefore, only differences in *β* due to delay would be notable. The only shift in *β* was from the immediate to the 20-minute delay for not-presented statements, $F(2, 58) = 9.71$, $p < .01$.

An unconfirmed prediction. The relative performance on the recognition and plausibility tasks shifts with delay as predicted by the schematic model in Figure 1. At a 2-day delay, statements that have been presented are judged plausible faster than they are recognized as having been presented. Similarly, subjects can judge implicit or unstated inferences as plausible faster than they can determine that such statements were not read. However, one prediction has not been confirmed: Subjects are not faster to judge an implicit statement as plausible than to recognize that statement when it has been presented in a story. Initially, subjects were over 1,300 msec slower to judge plausibility in this contrast, and the difference reduces to less than 50 msec, yet the predicted crossover did not occur. On the other hand, the speed–accuracy trade-off may account for the failure of this prediction.

The lack of a complete plausibility judgment advantage may be due to the high error rates in the recognition condition. Plausibility errors are on the order of 16%. In contrast, the error rate for recognition judgments is 40%. Moreover, this high rate of error is primarily due to bad performance in the not-stated condition, where subjects have a 60% false alarm rate. Presumably, subjects treat the recognition task as a plausibility judgment task and respond positively to plausible statements rather than search memory for the relevant fact. This interpretation is consistent with the false alarm rate of almost 70% for highly plausible statements and 50% for moderately plausible statements.

Given these error data and the "speed-accuracy trade-off" notion mentioned earlier, it seems unfair to conclude that it is less efficient to judge a not-presented statement as plausible than to recognize a presented statement. If subjects could be encouraged to be more accurate in the recognition task, or somewhat less accurate in the plausibility judgment task, the response times would reverse. This was the principal reason for conducting Experiment 2.

Experiment 2

Experiment 2 was basically the same as Experiment 1, with one essential modification: The accuracy and speed of each subject's performance was monitored on-line while he or she participated in the experiment. When too accurate the subject was told to speed up after slow, correct trials; when too inaccurate the subject was told to be more careful after error trials. The motivation of this manipulation was to shift the performance deficit from response accuracy to response speed and to equate accuracy across conditions to facilitate comparison of response times. This procedure was used successfully in Experiment 2 of Reder and Anderson (1980), where a similar problem had been encountered.

Method

Design and procedure. The general design and procedure were identical to those of Experiment 1, with the following exception: Subjects were told to modify their behavior when accuracy was above or below one standard deviation of 80% accuracy. The algorithm used was as follows: Where N is the number of questions that have been asked since the experiment began, when $.8N - .4\sqrt{N}$ was greater than the number of correct responses made thus far and on the current trial the subject made an error, then the terminal would display the following message, "Slow down. You are making too many errors"; when $.8N + .4\sqrt{N}$ was less than the total number of correct responses made thus far and on the current trial the subject's response was accurate and took longer than his average response time, the screen displayed the message, "You are responding too slowly. Please speed up."

Subjects. In the immediate delay condition there were 20 subjects in the recognition task and 22 in the plausibility judgment task. At the 20-minute delay, there were 14 subjects in the recognition task and 16 in the plausibility task. In the 2-day delay condition, there were 17 subjects in recognition task and 15 in the plausibility judgment task. As before, the task took approximately 25 minutes and subjects were given either money or course credit for participation.

Results and Discussion

Table 3 presents the data from Experiment 2, organized in a fashion similar to that in Table 1. Figure 4 displays the response time data for plausible statements collapsed over variations in plausibility.

Several analyses of variance (ANOVAS) were performed. One set (on both RT and accuracy) corresponds to those of Experiment 1 and the other analyses combined the data from Experiment 1 and 2. This combination resulted in a 2 × 2 × 2 × 2 × 3 ANOVA.

The pattern of significant results was the same as in Experiment 1 except that both delay and type of task now had effects on response time, $F(2, 98) = 4.68$, $p < .02$, and $F(1, 98) = 11.23$, $p < .01$, respectively. Subjects tend to take longer with delay in both the recognition and the plausibility tasks, but the increase is much greater for recognition. The interaction of delay with type of task was only marginally significant, $F(2, 98) = 2.9$, $p < .06$. However, the linear component to this interaction is significant, $F(1, 98) = 5.5$, $p < .025$, indicating that subjects slow down more in the recognition task.

An analysis that combines the data from the two experiments yields the same pattern, except that marginally significant effects are now quite significant, for example, the delay by task type interaction on RT, $F(2, 210) = 6.78$, $p < .01$.

As in Experiment 1, accuracy in the plausibility judgment task was much less affected by delay than in the recognition task, $F(1,\ 98) = 70.75$, $p < .01$. Subjects were now faster in all plausibility conditions than they were in Experiment 1. The biggest effect of accuracy monitoring was on the immediate not-stated condition (see Figure 4). Subjects were much faster (over 30% faster) to accept plausible statements that were not presented. This result is consistent with the notion that subjects had used a strategy of trying direct retrieval first in the immediate-delay condition in Experiment 1 but in Experiment 2 used plausibility judgment as a first strategy more often. For not-presented statements, trying direct retrieval first is a big disadvantage.

There are several other things to note. Subjects are faster to make plausibility judgments than recognition judgments in all conditions except in the immediate-delay condition, and subjects are faster overall in the plausibility judgment task, $F(1, 98) = 11.24$, $p < .01$. Monitoring accuracy caused subjects to respond slower in the recognition task at the 20-minute and 2-day delay but faster with immediate delay. Table 4 gives the d' and β values for Experiment 2.

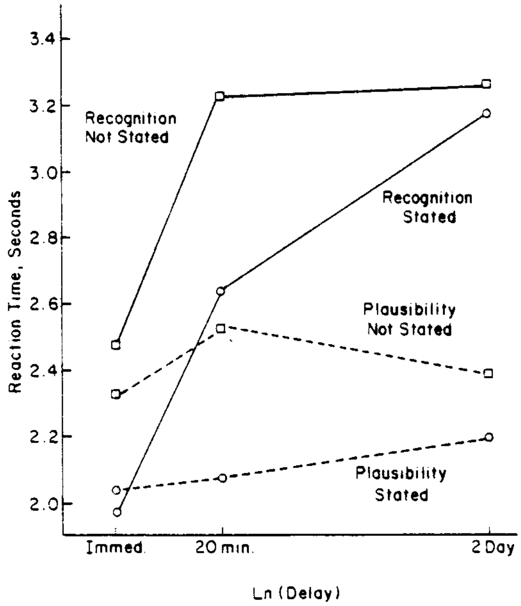


Figure 4. Mean reaction times in Experiment 2 for plausibility and recognition judgments as a function of whether the probe had been stated in the story, plotted across levels of delay.

The overall value of d' was unaffected by the monitoring instructions in the recognition task, $F(1,\ 103) = .84$ (compare Tables 2 and 4). Apparently, the information is no longer available after delays to allow for more accurate recognition judgments. The

Table 3
Mean Response Times (in seconds) and Error Rates for Experiment 2

	Judgment task			
	Recognition		Plausibility	
Delay	Stated	Not stated	Stated	Not stated
Immediate				
High plausibility	1.91 (.13)	2.51 (.36)	2.03 (.10)	2.17 (.13)
Medium plausibility	2.07 (.12)	2.46 (.17)	2.07 (.12)	2.50 (.28)
Implausible			2.20 (.15)	
20 minutes				
High plausibility	2.56 (.14)	3.15 (.53)	2.05 (.08)	2.36 (.13)
Medium plausibility	2.71 (.19)	3.25 (.29)	2.12 (.12)	2.68 (.25)
Implausible			2.59 (.14)	
2 days				
High plausibility	3.10 (.27)	3.11 (.58)	2.14 (.16)	2.27 (.16)
Medium plausibility	3.17 (.31)	3.35 (.43)	2.26 (.13)	2.51 (.28)
Implausible			2.24 (.15)	

Note. Error rates are in parentheses.

Table 4
d′ *and* β *Values for Experiment 2*

	Judgment task					
	Recognition		Plausibility stated		Plausibility not stated	
Delay	*d′*	β	*d′*	β	*d′*	β
Immediate						
Medium plausibility	2.36	1.30	2.52	1.32	1.87	2.01
High plausibility	1.66	.60	2.63	1.07	2.53	1.26
20 minutes						
Medium plausibility	1.58	1.04	2.36	1.00	1.86	1.57
High plausibility	1.04	.61	2.60	.81	2.32	1.02
2 days						
Medium plausibility	.78	1.01	2.22	.99	1.68	1.45
High plausibility	.56	.69	2.16	1.08	2.08	1.09

decline in d' with delay was quite significant for recognition in Experiment 2, $F(2, 47) = 25.6$, $p < .01$, as it has been in Experiment 1. The value of d' was affected by monitoring for plausibility subjects, $F(1, 108) = 8.9$, $p < .01$, because of the immediate condition.

One conclusion about the accuracy-monitoring manipulation is that because plausibility subjects can speed up in all conditions with, generally, no loss in accuracy, they were not performing optimally. The improvement in performance for plausibility subjects is most likely due to more emphasis on trying to use the plausibility judgment mechanisms than more emphasis on using fact retrieval because **the accuracy** in the not-stated conditions is helped more than in the stated conditions. This conclusion will be tested using the model fitting described below.

A second conclusion is that because recognition subjects slowed down with no improvement in accuracy, they cannot improve their performance because the memory traces are gone. All they can do is be more accurate in reporting what they know, by guessing less often (thereby making fewer false alarms) and saying "no" when they cannot find the statements (thereby missing more presented statements). Finally, the prediction was confirmed that subjects should be faster to judge a statement as plausible when it was not presented than to recognize it when it was presented.

Comparison of Model With Experimental Results

The model proposed earlier implies that both the direct retrieval strategy and the plausibility judgment strategy are used for both types of question-answering tasks and that the likelihood of selecting a particular strategy varies with the situation. The data from these experiments indicate that in both the plausibility task and the recognition task subjects are using both types of strategies to make their decisions. How much subjects rely on direct retrieval and how much on plausibility judgments depends on which judgment they are supposed to make, the test delay, and whether they are cajoled into being more or less accurate. Subjects use the plausibility judgment process to some extent in the recognition task because they are faster to verify highly plausible statements than moderately plausible statements. This difference grows with delay, suggesting that even recognition subjects are using plausibility more often to answer questions at longer delays.

The fact that explicit inferences are judged as plausible faster than are implicit ones is evidence that plausibility subjects sometimes adopt the retrieval strategy. Other evidence is the fact that the biggest plausibility effect obtains in the not-presented plausibility task condition where direct retrieval cannot be used. The advantage of stated probes, however, declines from the immediate to 2-day delay, suggesting that subjects put less emphasis on direct retrieval at longer delays.[4] The fact that monitoring performance in the plausibility judgment condition had the intended effect of speeding up responses with little reduction in accuracy is consistent with the idea that subjects can simply use direct retrieval less often and make plausibility judgments more often. This strategy shift will produce a speedup with no deleterious effects on accuracy.

Fitting the Model to the Data

The equations given earlier, when the strategy selection model was formalized, must be modified slightly to fit the data. For example, the search times, *S1* and *S2*, and the judgment times, *J1* and *J2*, for the two strategies cannot all be separately estimated. Instead *A* is estimated for the recognition process, representing both time to retrieve a proposition, *S2*, and judge its adequacy, *J2*. *G* represents the time to glean enough relevant information, *S1*, and compute whether the probe is plausible on the basis of this material, *J1*. (Table 5 lists the names and descriptions of all parameters used.) Figure 5 is a modified graph of Figure 2. At the terminus are the values that were actually fitted to the data.

The assumption mentioned earlier that time to retrieve a specific fact increases with delay dictates that three separate estimates for *A* should be derived, corresponding to the three different delays used in the experiment. The decision time for judging the adequacy of a retrieved fact, *J2*, should not vary with delay, hence, any change in *A* should reflect a change in search time, *S2*, for direct retrieval.

In contrast, *S1* is not expected to vary appreciably with delay because of the large redundancy in acceptable facts caused by elaboration. Because many different subsets of facts will satisfy the search stage of plausibility, this process has an advantage over recognition in being relatively impervious to delay manipulations. The time to compute plausibility, on the other hand, is expected to be affected by the inherent plausibility of the probe. Therefore, different values of *G* are estimated for highly and moderately plausible statements. A moderately plausible statement is assumed to be more plausible if it is in fact stated in the story. (As noted before, the determination of plausibility was done when the items were not explicitly presented.) Presumably, making a statement that is already considered highly plausible explicit has little effect on its plausibility. Hence three levels of *G* were fit for an experiment: highly plausible, moderately plausible stated, and moderately plausible not stated.

Other variables given in the earlier equations can be estimated from the data. The variable α—the probability of finding a statement implausible—may assume three different values depending on the inherent plausibility of the statement: α for highly plausible, α' for moderately plausible statements that were explicitly presented, and α'' for moderately plausible statements that had not been presented.

The values for *f*, the probability of finding the fact in memory at different delays when searching for a specific fact, are estimated indirectly, as are the values for *A*, the time to activate the closest match to the test probe. The estimates of *f* and *A* involve the assumptions that (a) search for a specific fact terminates when a candidate proposition is activated that satisfies some criterion of closeness to the probe or by a cutoff in search time, C, is reached and (b) the distribution of times to activate a specific proposition and the distribution of cutoff times are exponential. The time for the candidate to be activated, if there were no cutoff, would be R_i (i representing level of delay). This notion of a race between competing processes with exponential distributions (in this case R and C) has been used before (e.g., King & Anderson, 1976; Mohs, Westcourt, & Atkinson, 1975). The expected activation time for the retrieved proposition (activated prior to the cutoff) is

[4] The reason the difference between stated and not stated remains for moderately plausible statements is due to the fact that presenting a moderately plausible inference as part of a story necessarily increases its plausibility. In other words, it is not an advantage due to being "presented" per se but due to becoming more plausible.

$$A_i = \frac{R_iC}{R_i + C}. \qquad (10)$$

The probability of finding the fact prior to the cutoff is

$$f_i = \frac{C}{R_i + C}. \qquad (11)$$

If a candidate is retrieved prior to the cutoff, the probability of accepting it is f when the probe has been presented in the story, but a fact may be erroneously accepted, with probability γ, when the probe was not presented.

If the probe is not found in memory when direct retrieval was tried, then with probability y the subject quits and with $1 - y$ the subject tries to answer the question using

Table 5
Parameters of the Model and Best Estimates in Experiments 1 and 2

		Experiment	
Description	Parameter	1	2
Probability of trying to judge plausibility first			
Immediate, recognition task	x_1	.21	0
20 minutes, recognition task	x_2	.51	.26
2 days, recognition task	x_3	.73	.25
Immediate, plausibility task	x'_1	.57	.89
20 minutes, plausibility task	x'_2	.93	.87
2 days, plausibility task	x'_3	.99	.91
Probability of not trying plausibility after search fails in recognition task	y_1	.99	.63
Probability of not going on to try plausibility if search fails in the plausibility task	y_2	0	0
Probability of retrieving exact fact sought[a]			
Immediate	f_1	.80	.78
20 minutes	f_2	.79	.69
2 days	f_3	.75	.66
Probability of "finding fact" when not presented	γ	0	0
Probability of deciding that the statement is implausible			
Highly plausible	α	.10	.12
Moderately plausible when explicit in story	α'	.16	.14
Moderately plausible when not stated	α''	.28	.30
Time to complete search stage (assuming cutoff)[b]			
Immediate	A_1	2.42	1.88
20 minutes	A_2	2.51	2.68
2 days	A_3	2.96	2.99
Time to evaluate plausibility of statement			
Highly plausible	G	2.41	1.98
Moderately plausible, stated	G'	2.64	2.07
Moderately plausible, not stated	G''	2.89	2.28
Negation time constant	K	.17	.28
Cutoff time on search	C	12.06	8.69
Time to retrieve or activate a given proposition, assuming no cutoff			
Immediate	R_1	3.02	2.40
20 minutes	R_2	3.17	3.87
2 days	R_3	3.92	4.55

[a] See Equation 11.
[b] See Equation 10.

plausibility judgments. The value of y is assumed to vary depending on whether the subject was asked to recognize or judge plausibility. In recognition, the probability of quitting is y_1; in plausibility it is y_2. Presumably y_2 should be very small in that if direct retrieval were tried first and failed, the person would then try to evaluate the plausibility of the statement given the task demands. (See Table 5 for a description of each parameter.)

The same general model should apply when a subject's speed and accuracy of performance are monitored. The time constants and the probability values for most conditions should be the same as in the unmonitored task, with a few exceptions: Depending on accuracy level and the resulting feedback, the probability of selecting the direct retrieval or plausibility judgment mechanisms may shift. Also, when recognition fails the likelihood of going on to plausibility may be less in the monitored task. Finally, the cutoff on letting activation spread might be reduced.

The formula for a specific prediction can be constructed using Figure 5 (or the equations given earlier) and the appropriate parameters for a given condition. For example, the expected accuracy (percent correct [PC]) judge plausibility for a moderately plausible statement that was not presented, and that is tested immediately after reading the story, would be

$$E(PC) = x'_1(1 - \alpha'') + (1 - x'_1)[\gamma + (1 - \gamma)(1 - y_2)(1 - \alpha'')]. \quad (12)$$

The reaction time equation for the same condition is

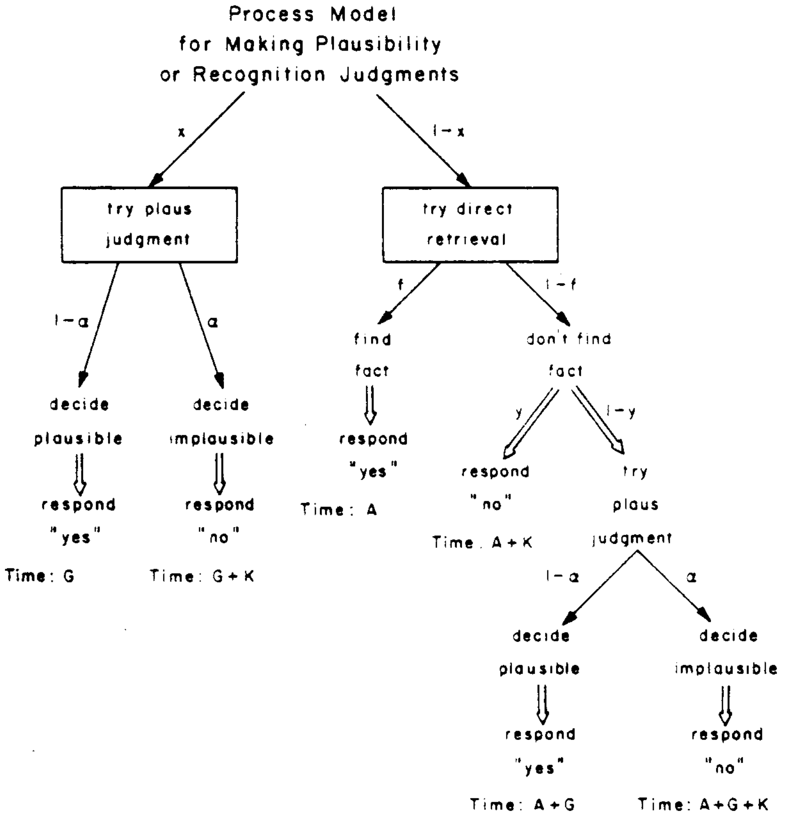


Figure 5. Revised model of alternative strategies reflecting the parameters than can be estimated or derived.

$$\text{E(RT/“yes”)} = \frac{x'_1(1 - \alpha'')G'' + (1 - x'_1)[\gamma A_1 + (1 - \gamma)(1 - y_2)(1 - \alpha'')(A_1 + G'')]}{\text{PC for this condition}}. \quad (13)$$

For a different delay condition, the value of A_i will change from A_1 to A_2 or A_3 as will x'_1. For highly plausible statements, α'' and G'' would be replaced with α and G, respectively.

For judging the plausibility of a statement that is highly plausible and had been stated in the story, 2 days after reading the material, the following equations would be used to predict accuracy and response time:

$$\text{E(PC)} = x'_3(1 - \alpha) + (1 - x'_3) \times [f_3 + (1 - f_3)(1 - y_2)(1 - \alpha)] \quad (14)$$

$$\text{E(RT/“yes”)} = \frac{x'_3(1 - \alpha)G + (1 - x'_3)[f_3 A_3 + (1 - f_3)(1 - y_2)(1 - \alpha)(A_3 + G)]}{\text{PC from Equation 14}}. \quad (15)$$

These equations are used for any presented statement in the plausibility judgment task, modified for values of f_i and A_i and x'_i at different delays; α becomes α' and G becomes G' for moderately plausible statements.

In the recognition task the equations are the same as for plausibility when the statement had been explicitly presented. The only difference is that x_i replaces x'_i and y_1 replaces y_2.

However, for conditions where the probe had not been presented, the equations are not similar to the plausibility conditions because the correct response is different. For example, the 20-minute delay condition for highly plausible statements is represented by the following equations:

$$\text{E(PC)} = x_2\alpha + (1 - x_2)(1 - \gamma) \times [y_1 + (1 - y_1)\alpha] \quad (16)$$

$$\text{E(RT/“no”)} = \frac{x_2\alpha(G + \text{K}) + (1 - x_2) \times (1 - \gamma)[y_1(A_2 + \text{K}) + (1 - y_1)\alpha(A_2 + G + \text{K})]}{\text{PC from Equation 16}}. \quad (17)$$

Because this condition is for not-presented statements, α and G would be replaced by α'' and G'', respectively, for moderately plausible statements.

Estimating the Parameters

Given the constraints of the model described above, there are 20 free parameters to estimate from Experiment 1 or 2; the x parameters for the plausibility and recognition tasks for the three levels of delay, which make six. There are two ys, ~~and~~ one for recognition and one for plausibility. The three values of f and of A are estimated from the three values of R and the cutoff C. The time to make a guess, G, varies with whether the statement is highly plausible or moderately plausible or was predetermined to be moderately plausible but was presented in the story. So there are three Gs as there are three values of α, the probability of finding a probe plausible. In addition to the cutoff, C, there is a constant, K, for negative response times and γ. (See Table 5.)

In each experiment, collapsing over subjects, there are 48 data points to fit (for 24 conditions with two dependent measures—RT and accuracy). The STEPIT program (Chandler, Note 1) was used in order to find the parameter values that would give the best fit to the data, by minimizing the value of a chi-square-like statistic, denoted C^2. This statistic had 28 degrees of freedom (48 point, 20 values). The formula used to minimize C^2 was

$$C^2 = \sum_{i=1}^{24} \left[\left(\frac{\hat{\text{RT}}_i - \overline{\text{RT}}}{s_{\overline{\text{RT}}}} \right)^2 + \left(\frac{\hat{\text{PC}}_i - \overline{\text{PC}}_i}{s_{\overline{\text{PC}}}} \right)^2 \right], \quad (18)$$

where i indexes the 24 conditions (Task × Probe Plausibility × Stated/Not Stated × Delay), $\overline{\text{RT}}$ means observed response time, RT means *predicted*, and s corresponds to

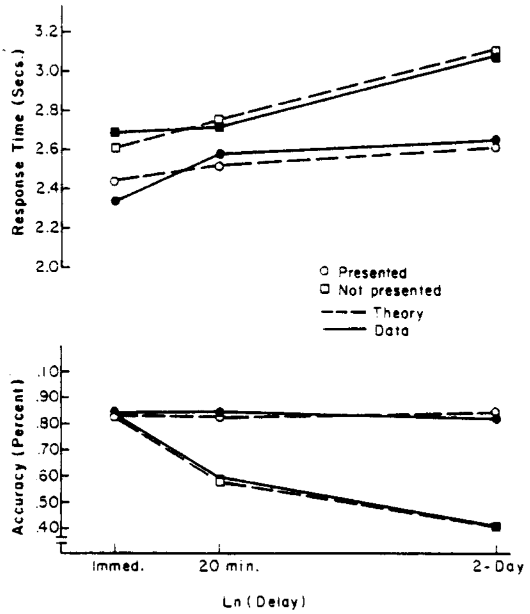


Figure 6. Comparison of theoretical and empirical data points for both reaction time and percentage correct for recognition subjects in Experiment 1 (averaged over plausibility of the test probe).

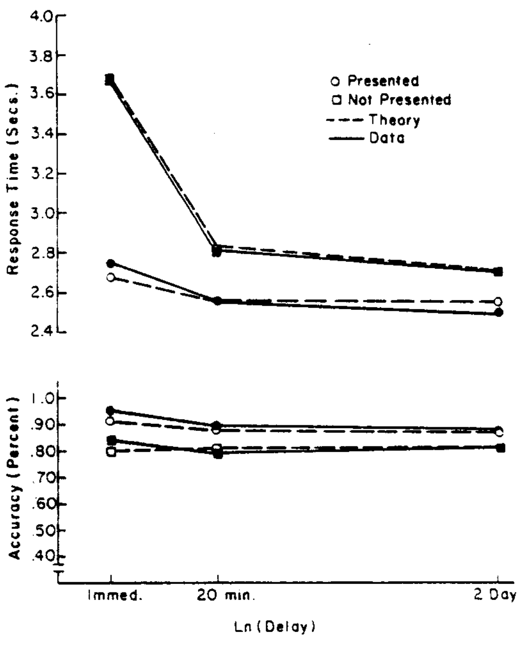


Figure 7. Comparison of theoretical and empirical data points for both reaction time and percentage correct for plausibility judgment subjects in Experiment 1 (averaged over plausibility of the test probe).

the standard error for a given measure in each of the delay conditions. A similar formula was used in model fitting by King and Anderson (1976) and Anderson (1981).[5]

The least C^2 value obtained by the model was 68.48 in Experiment 1 and 96.77 in Experiment 2. Both values indicate significant ($p < .01$) deviations of the observed from the predicted values. However, another index of goodness of fit—the proportion of variance accounted for, obtained by squaring the coefficient of correlation between observed and predicted data points—seems quite good. This r^2 value calculated over both experiments is .96 for reaction time and .90 for accuracy. On the other hand, to claim to account for 96% of the variance among reaction times means may be misleading in that the more parameters in a model, the easier it is to obtain a good fit.

A procedure used by Reed (1976) adjusts the value of r^2 to reflect the number of free parameters. This is done by altering the original variance fraction to divide the sums of squares in numerator and denominator by their corresponding degrees of freedom, as given in Equation 19.

$$r^2 = 1 - \frac{\sum (x_i - \hat{x}_i)^2/(h - k)}{\sum (x_i - \bar{X})^2/(h - 1)}, \qquad (19)$$

where h is the number of empirical points x_i, k is the number of free parameters in the theoretical function, $\hat{x}_i$ are the theoretical values corresponding to x_i, and $\bar{X}$/is the grand mean of x_i. A problem arises in applying this formula because of the need to calculate separate correlations for reaction times and for accuracy. The parameters were estimated simultaneously, and it is not clear how to assign the degrees of freedom used for estimation to the two dependent measures. As an ad hoc assumption, the 40 parameters in the two experiments were split into 20 parameters for each response type and each r^2 was adjusted, assuming that h = 48 and k = 20. The resulting r^2 values were .93 for reaction time and .83 for accuracy.

[5] Miller and Greeno (1978) compared the parameter estimates and the chi-square-like value obtained by King and Anderson (1976) with their own and they found them indistinguishable.

Even with these corrections the variance fraction seems quite good.

Another way to appreciate the goodness of fit is by inspecting Figures 6 through 9, which show how close the obtained data points are to the theoretic RT and accuracy functions. Both the fit for RT and percentage correct seem quite close for both experiments. Critically examining the parameter estimates from Experiments 1 and 2 provides another way to evaluate the "goodness" of the model.

Evaluation of the Parameter Estimates

A number of parameters in Experiments 1 and 2 are not expected to vary according to the model. For example, α, the probability of deciding that a statement is implausible, should not vary as a function of whether the subject's accuracy was monitored. On the other hand, x, the probability of trying plausibility rather than direct retrieval can very well change when subjects are asked to try to be more accurate. Because the data from the two experiments were fit separately, the parameter estimates can be compared to see if they are close for those that are theoretically expected to be close. Also, the values themselves can be examined to see if they are reasonable, for example, do they vary appropriately with delay, task, and plausibility?

The final estimate for each parameter is given in Table 5. The model expects nonstrategy parameters to have the same value across experiments namely, f, γ, α, G, K, and A. "Strategy" variables (the x and y parameters), on the other hand, would be expected to change because they reflect choices that, at some level, are under the subject's control. The x variables represent the probability of using the plausibility strategy first. This probability would change if subjects' accuracy were monitored. Similarly, y, the probability of trying plausibility when direct retrieval fails, should vary. Because the program that fit the data was not aware of these theoretical constraints, we can see whether the parameters that should be the same across experiments correlate more closely than those that should not. The nonstrategy parameters correlate well: r =

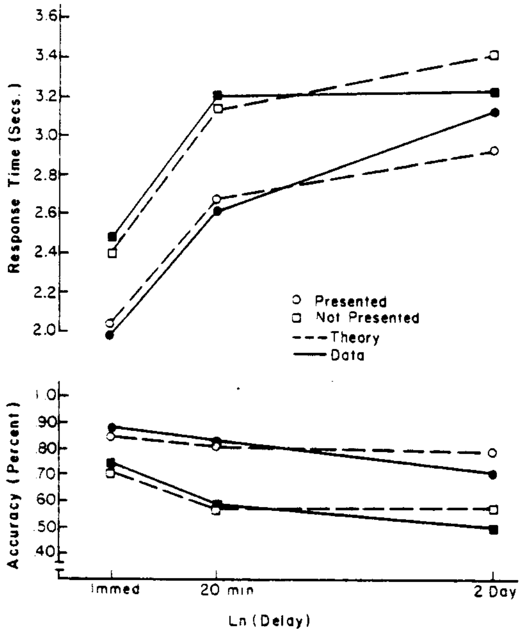


Figure 8. Comparison of theoretical and empirical data points for both reaction time and percentage correct for recognition subjects in Experiment 2 (averaged over plausibility of the test probe).

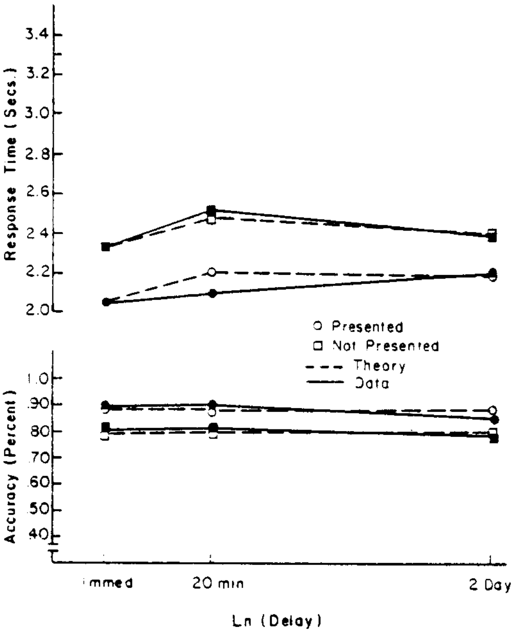


Figure 9. Comparison of theoretical and empirical data points for both reaction time and percentage correct for plausibility judgment subjects in Experiment 2 (averaged over plausibility of the test probe).

.99 for the probabilities and $r = .93$ for the times; the strategy parameters have a correlation of .79.

Another test of the model's credibility is to see if the values of the parameters vary as the model and task characteristics would expect. First consider the strategy parameters expected to vary with accuracy monitoring, that is, the x and y variables. In Experiment 1 the probability of using plausibility to make a judgment without first trying direct retrieval increases with delay, presumably because the incidence of forgetting increases. This is true for the recognition task (x) and the plausibility judgment task (x'). At a 2-day delay, recognition subjects are plausibility as the primary strategy 73% of the time (estimated by model fit; see Table 5). Of course, those asked to judge plausibility are even more inclined to use that strategy than are recognition subjects so task directions do affect the strategy selected. Nonetheless, in the immediate condition subjects asked to judge plausibility are almost as likely to try direct retrieval first as they are plausibility, $x'_1 = .57$, presumably because the lexical traces are still strong. Undoubtedly, when lexical information is highly available, direct retrieval is an appealing strategy. On the other hand, plausibility subjects try direct retrieval first only 7% of the time at the 20-minute delay and only 1% of the time at the 2-day delay. These parameter estimates explain the big speedup in the not-presented condition of the plausibility task from the immediate to the 20-minute delay. If direct retrieval is tried first, it will fail because the probe had not been stated in the story and then the plausibility strategy would have to be tried.

The values of x vary appropriately with delay and task demands in Experiment 2. The shift in values caused by accuracy monitoring also seems sensible. Again, recognition subjects increase their reliance on the plausibility strategy as a first option at longer delays. However, the stress on accuracy makes them much less likely to try this first than are subjects in the unmonitored experiment. In the immediate condition, when memory traces are strong and retrieval of a fact is relatively fast, subjects apparently never bothered to try to judge plausibility first. The reason that using the plausibility strategy, before trying recognition, dropped for recognition subjects in general is due to the realization that if they used plausibility to make their judgments, they would make more errors. When they failed to find the fact, however, they were more inclined to try plausibility than they were in Experiment 1 (y_1 in Experiment 2 is less than in Experiment 1). Perhaps this reflects the realization that at a delay memory traces are weak. Given that finding was more accurate, subjects were willing to continue to work at deciding whether a statement was presented.

The probability of "going down" the plausibility branch first did not change as much in Experiment 2 for subjects in the plausibility condition. The only major shift was in the immediate condition where the rate went up from 57% to 89%. This might have been caused by the fact that subjects were very accurate in this condition, causing the program to tell them to speed up on a slow, accurate trial. This suggests that subjects considered judging plausibility as a faster but less accurate decision mechanism. Judging plausibility first is indeed a less accurate mechanism even for a plausibility task, given that subjects asked to do this who first tried direct retrieval always used the plausibility strategy when the statement was not found. That is, $y_2 = 0$ for both experiments. Unlike subjects who were asked to make recognition judgments, plausibility subjects felt that they could not say "no" (implausible) simply because they could not retrieve a statement. Hence, the model could be simplified by deleting parameter y_2.

Now consider the remaining probability parameters. When direct retrieval is tried, the probability of finding the fact, f, should decrease with delay, and it does.[6] The probability of "finding" a fact that was not presented, γ, is estimated at zero in both experiments. This seems quite sensible and allows a second simplification of the model,

[6] The parameters f and A will be discussed rather than R and C from which they were derived because the former are directly involved in the predictions.

namely, that nonstudied facts are never retrieved (reducing the total number of estimated parameters to 18).

When the plausibility judgment strategy is used, the probability of deciding incorrectly that a statement is implausible, α, increases the less plausible the statement is (highly plausible statements have the smallest estimates of α and not-stated moderately plausible statements have the largest value). The probability estimates for α are quite similar across experiments.

The estimates for the time parameters also seem sensible. When trying direct retrieval the time to complete the search stage, A, increases with delay as expected and is fairly close in value across the two experiments. Subjects take less time to evaluate a statement's plausibility, G, when the statement seems more plausible. The pattern is constant across experiments, although subjects seem to make their judgments somewhat faster in Experiment 2. It is noteworthy that the estimated retrieval times are faster than the plausibility times in the immediate condition but longer in the 20-minute and 2-day delay conditions.

Fitting a Competitive Model

To appreciate the appropriateness of the proposed strategy-selection model and the insufficiency of an obvious competitor, both models should be fit to the data. I will call the competitor examined the *default* model because most theorists at least implicitly assume it when discussing question answering. This model can be represented by a subset of Figures 2 or 5. The top branch of that tree structure allows a choice in first strategy, namely, selecting direct retrieval first or judging plausibility first. The default model posits that people always try to retrieve the queried fact from memory first. Therefore the top branches, with the probabilities x and $1 - x$ can be deleted, along with the entire left side of the tree. What is left is the strategy of trying direct retrieval first and, should that fail, then the possibility of trying to judge plausibility.

Equations are derived for this default model in a manner analogous to that used for the strategy-selection model proposed here. The only difference in any equation is that all x_i and $x_i' = 0$. Compare the equation below (Eq. 20) for time to recognize a presented statement that is highly plausible in the immediate condition with the comparable equation (Eq. 15) of the strategy-selection model, replacing y_1 for y_2.

$$E(\text{RT}/\text{``yes''}) = \frac{f_1 A_1 + (1 - f_1)(1 - y_1)(1 - \alpha)(A_1 + G)}{f_1 + (1 - f_1)(1 - y_1)(1 - \alpha)} \qquad (20)$$

The formula for accuracy is just the denominator in Equation 20.

The data were fit separately for Experiments 1 and 2, estimating 14 parameters in each fit, again using STEPIT (Chandler, Note 1). The C^2 statistic that the program attempted to minimize was the same as before (Eq. 18) and had 34 degrees of freedom (48 points, 14 values). The minimal C^2 statistic was 489.71 for Experiment 1 and 333.58 for Experiment 2. These values, of course, suggest significant deviations ($p < .01$) of the observed from predicted values. (The strategy-selection model's C^2 values were 68.5 and 96.8, respectively). The default model accounts for 62% of the variance among RT means and 64% of the variance among accuracy means, using an uncorrected r^2. Adjusting r^2 for the 14 free parameters, using Equation 19, the values are .47 and .50 for RT and accuracy, respectively. The adjusted r^2 values from the selection model, .93 and .83, respectively, compare quite favorably. Clearly the strategy-selection model provides a much better fit.

By comparing the predicted data points to the observed points and inspecting the derived parameter values, it becomes clear why the more complex model is needed. Plausibility judgments are often faster than recognition judgments in the data; however, this model requires that direct retrieval be tried first. There is no way for the default model to find a set of parameters that allows plausibility to be faster than recognition.

In addition to the above problems, the default model is inadequate because it cannot account for why subjects are initially slower in the plausibility task than in the recognition task but are later faster for plausibility. It also cannot explain why there is a speedup in the plausibility task for presented statements.

Summary of Fit of the Model

Using several different criteria, the strategy-selection model provides a good fit to the data.

1. The fit is much better than that of its competitor, the default model.
2. The parameter estimates derived by the model using STEPIT (Chandler, Note 1) have sensible values, for example, probabilities of finding something implausible vary appropriately with judged plausibility, and forgetting probabilities vary appropriately with delay.
3. The parameters that should have been shared by Experiments 1 and 2 have estimates that are highly correlated and close in value, even though they were fit independently.
4. The values for the probability of selecting one strategy or the other make sense given the task asked of subjects, the strength of the memory traces, and whether accuracy is monitored.

General Discussion

Further Empirical Support for the Strategy-Selection Model

The fan effect: Resolving the paradox of interference. Previous research (e.g., Anderson, 1974, 1976; Lewis & Anderson, 1976; Thorndyke & Bower, 1974) has shown that the more facts committed to memory about a particular concept, the slower a person is to recognize or reject (as not studied) any statement related to that concept. The theoretical argument used to explain that phenomenon dubbed the "fan effect" (Anderson, 1974) need not concern us here (although it was mentioned briefly in the introduction). The apparent paradox discussed by Smith, Adams, and Schorr (1978) is relevant, however. They noted that the more facts an individual knows about a concept (the more expertise), the more difficult it is to answer a question about that knowledge. Smith et al. (1978) showed that the fan effect is greatly attenuated if the facts associated with a concept are thematically related. For example, if the concept were a fictitious individual (e.g., Marty), the fan effect would be much greater if the facts learned about Marty seemed unrelated to one another than if they were all related by some theme (e.g, christening a ship).

The model proposed here is relevant because the facilitation by related facts may be due to the opportunity to use plausibility as a judgment strategy even though the ostensive task is recognition. More facts about Marty will not slow response time if any fact about Marty christening a ship can be used to "recognize" any other fact about Marty christening a ship. The foils used by Smith et al. (1978) did not preclude this strategy.

Experiments by Reder and Anderson (1980) supported this view: In blocks of trials where the targets were tested in the presence of thematically related foils (a fact about ship christening not studied with Marty), the fan effect was as large as with unrelated facts; in blocks of trials where the foils were unrelated and did not preclude judging plausibility, the fan effect was diminished.

Because the fan effect was diminished by an apparent shift in strategy from judging recognition to judging plausibility when the ostensive task was recognition, it seemed reasonable to assume that the effect of fan would be even smaller (and perhaps negative) if subjects were actually asked to judge plausibility. That is, the weak fan effect in the presence of unrelated foils may be a mixture of two strategies: judging recognition as requested, which produces a strong fan effect, and sometimes judging plausibility, which may generate a negative fan effect. There is reason to expect a negative function or facilitation of fan in the plausibility judgment task. The argument is related to that given with Figure 1. The more facts that are available from which to search memory, the faster a relevant subset can be retrieved for the decision stage.

An experiment by Reder and Ross (Note 2) is consistent with this notion. We asked subjects to recognize statements in some blocks of trials and, in other blocks, asked them to decide if statements were consistent with studied statements. The data replicated those of Reder and Anderson (1980) for the two types of recognition, that is, a larger fan effect in the presence of related foils. In the consistency judgment block, negative fan functions were obtained—subjects took less time to judge that a fact was consistent with information studied, the more facts they had studied on that topic. The negative fan function was larger for statements that could not be judged using direct retrieval, that is, those statements that had not been studied with the probed character. The fact that the negative fan effect was smaller for studied sentences indicates that subjects were sometimes still making their decisions by direct retrieval. The fact that there was a negative fan effect for these sentences, rather than the positive fan effect obtained in recognition blocks, indicates that plausibility was chosen more often than before.

Manipulations of strategy selection other than task and delay. Recall that in Experiment 1 there had been an enormous speedup on not-stated items for the plausibility subjects from the immediate condition to the 20-minute condition. The explanation given was that subjects were quite prone to try direct retrieval first at the shortest delay, which slowed them down when the fact had not been presented. When memory traces weakened over time, there was a shift in the distribution to make plausibility judgments without first trying direct retrieval; this resulted in a speedup in the mean reaction time of the not-stated items in the plausibility task. This same shift in strategy should be possible without using a longer delay. If subjects have no verbatim traces, then they would not be tempted to use direct retrieval. Recently, I have performed another experiment (Reder, Note 3) that is intended to test this aspect of the model, that is, shift subjects' strategies without manipulating delay or monitoring.

The new follow-up experiment was essentially a replication of the immediate condition of Experiment 1, with one exception: The plausibility subjects never saw any of the test items presented in the story. For comparison, recognition subjects were run with materials identical to those used before. The results support the strategy-selection model and the interpretation of earlier results: The data for the recognition subjects were quite comparable to those from Experiment 1. In contrast, for the plausibility task the moderately plausible, not-presented RT was now 2.48 sec rather than 4.04 sec, ~~respectively,~~ and the highly plausible was now 2.34 sec rather than 3.29 sec, ~~respectively,~~ an average speedup of 1.25 sec (accuracy also improved for the plausibility condition). The response times for plausibility judgments in the immediate not-presented condition looked like the 20-min delay condition, except that the current subjects were slightly faster than ~~the~~ 20-min condition ~~of Experiment 1 as well~~. This is to be expected because subjects have very little reason ever to try direct retrieval given that the test items were never presented.

Comparison of Model Predictions With Other Results in the Literature

The work of Singer (1979a, 1979b) is quite relevant to my own work. He investigated whether the inferences that often cause false alarms, as in the Bransford and Franks (1971) task, are drawn when the passage is read or during the test. He found, as I have (Reder, 1976, 1979), that it takes subjects longer to respond to implicit statements (not-presented inferences) than to explicit statements and thereby concluded that inferences are for the most part not drawn during comprehension.

Singer did not discuss the efficiency issue of fact retrieval versus plausibility (inferential) judgment, although his data (Singer, 1979a) allow such contrasts. With a 20-minute delay between study and test, his subjects were faster to verify an explicit (presented) statement than to recognize it, and truth judgments were faster for implicit statements than were recognition judgments for these not-presented statements. He did not find implicit verification to be faster than explicit recognition, but with a longer delay, or better control of high error rates, I think Singer's experiments would have shown that pattern too.

Singer's conclusion that subjects do not infer during comprehension but at testing is in direct contrast to a statement by Graesser, Robertson, and Anderson (1981): "The inferences generated from the question-answering procedure were likely to have been constructed during comprehension, as opposed to being fabrications that are invented during questioning" (p. 2). In some sense both positions are correct. Readers generate many inferences and embellishments during comprehension; however, it is not likely that the exact statement queried is inferred prior to testing. Given that it is faster to compute a statement's plausibility at testing than it is to retrieve it, it is not surprising that people do not strive to infer absolutely everything during comprehension.

Given the position that people do not often use direct retrieval, one might wonder why explicit (presented) inferences are faster than implicit inferences. There are several reasons. First, in a search for relevant information, occasionally an exact match will be found and the computation stage can be skipped. Second, explicitly stating a fact causes it to be more plausible, making Stage 2 faster. A third reason is that stating the fact causes even more relevant elaborations to be generated, which affects the speed of Stage 1. In support of the last reason, Reder (1979) found that primed inferences were judged faster than were implicit ones. It was argued there that priming (asking related questions that omitted mention of the critical information) caused generation of more embellishment than does presenting the statement (compared with the not-presented condition). There will always be more relevant information in the explicit case than the implicit case, which means that search will usually be faster regardless of whether the exact fact is retrieved.

Recent studies by Camp (1978) and Camp, Lachman, and Lachman (1980) superficially seem in contradiction to my results. They found that *inferential questions* were verified more slowly than *direct access questions*. (Questions were paired with a single alternative.) Direct access questions were ones subjects could not typically recall but could recognize with high confidence. Inferential questions were designed so that most people would not have encountered them directly, but they could be answered (theoretically) by combining typically known facts. Examples of inferential questions used were: "Which way does the Statue of Liberty face?" (answer: southeast) and "Which horror-movie character might starve in Northern Sweden in the summertime?" (answer: Dracula). Examples of direct access questions were: "What is Bob Dylan's real name?" (answer: Zimmerman) and "What was the name of the U.S. freighter seized in 1975 by Cambodians?" (answer: Mayaguez). They found that questions whose answers intuitively seemed to require inference took longer to verify than questions whose answers were thought not to require inference.

This result is not inconsistent with the model proposed here for several reasons. First, they have shown that some questions are more difficult to answer than others (just as the present studies found moderately plausible ones more difficult). Second, the reason given earlier for why judging plausibility can be faster was based on the notion that the search stage can be faster for judging plausibility because a large number of relevant facts are available from which to select. However, in their experiment the inferential problems required that several specific facts be retrieved in order to answer a question. Those needed facts were often esoteric or weakly encoded in memory, making the search stage quite long for judging inferences. (For example, the information needed to determine the direction that the Statue of Liberty faces is quite weak in my memory, if present at all.) Third, the direct access task also differed from what I consider to be real direct access. Subjects were not required to discriminate a presented statement from a paraphrased one but could use reconstructive processes to discriminate target from foil. In summary, Camp et al. (1980) were comparing question answering for different types of material. The research reported here compares different types of question answering for the same type of material.

The introduction described a study by Kintsch (1974, Chapter 8), with McKoon and Keenan, that showed no significant difference in judgment time between explicit and implicit statements at a delay. In that experiment all subjects were making plausibility judgments. The most comparable contrast to their explicit versus implicit conditions in the current studies would be the presented versus not-presented statements for the plausibility judgments in Experiment 1. The highly plausible statements are most similar to their statements. The difference between the presented and not-presented probes is as large as theirs in the immediate task and as small as theirs at a delay. We both attribute the loss of advantage in the presented case to fading lexical traces. They take the negligible difference to mean that the inference is stored and retrieved directly in both the explicit and implicit conditions. However, the current theory suggests an entirely different explanation. Subjects might have been making their decisions by plausibility judgments. This alternate explanation is supported by the plausibility effects shown in my experiments.

An experiment by Baggett (1975), using a series of cartoons rather than a verbal story, is supportive of the positions argued here. Subjects viewed cartoons depicting a vignette such as going to a barber for a haircut. Subjects were asked to judge the consistency of verbal statements with respect to the vignette. Baggett manipulated delay of test and whether the information queried had been explicitly depicted in a cartoon or not. Note that in this task there can be no lexical trace and the encoding of the pictures will not necessarily match the test probes. In this case there was no effect of the explicit/implicit factor or the delay factor. Presumably, subjects were using solely the plausibility judgment mechanisms in all conditions.

A number of results in the literature can be reinterpreted using the notion of subjects' preferences for plausibility judgments. For example, the work of Bransford and Franks (1971), Bransford, Barclay, and Franks (1972), and Johnson, Bransford, and Solomon (1973) have all shown that test sentences implied by the studied material are "recognized" almost as frequently as are statements actually studied. Bransford et al. (1972) concluded that we store the gist, which includes inferences. At testing we retrieve a match to the test probe and cannot distinguish fact from inference. A different explanation involving the position presented here is that at testing we prefer to judge plausibility if the verbatim traces are weak. In the case of the Bransford et al. (1972) task, the traces are weak due to an extraordinary amount of lexical interference, that is, using the same phrases multiple times in the same passage. Therefore, these data may reflect what is happening at time of test more than what happened at time of study, that is, the inferences may not have been stored but rather computed at testing.

A common result in the text-processing literature (e.g., Mandler & Johnson, 1977; Meyer, 1975; Rumelhart, 1975; Thorndyke, 1977) is that information represented higher in the tree representation of the passage is better recalled. This result can be explained by the fact that central propositions are more plausible, without assuming that the statements are actually better being remembered. The higher level ideas are implied by the lower ideas and tend to be embellished more than are lower level ideas. These differences will make central ideas easier to reconstruct and to verify.

One of the most intriguing findings in the social psychological literature is the effect of order of presentation of attributes on judgments of personality (e.g., Asch, 1946) and on recall of those attributes (e.g., Anderson & Hubert, 1963; Dreben, Fiske, & Hastie, 1979). The basic finding that numerous theorists have modeled is that the first attributes described about a person have the greatest effect in subsequent impression-formation judgments, whereas the attributes best recalled are the ones presented later. Probably the attributes presented first more greatly affect the type of elaborations generated about a person than do the subsequent attributes. That is, people interpret and embellish subsequent information in light of the initial impression. Thus, it is easy to see why questions about impressions are affected more by the first facts even though the later facts are better recalled: Impression judgments are essentially plausibility judgments and will be affected by the elaborations as well as the studied material. Indeed, research by Keenan (Keenan & Baillett, 1980; Kennan, Note 4) is consistent with the view that trait judgments are plausibility judgments: the more information relevant to the trait that is known, the faster the judgment is made.

Implications for Other Tasks

The notion that plausibility judgments are often easier to make than recognition judgments can probably be extended to recall measures. Reconstruction is probably easier than direct recall, except at short delays. That is, Bartlett's (1932) notion that much of memory involves reconstruction is analogous to the claim that much of verification involves plausibility judgments. If the paradigm could be developed, I speculate that reconstruction would be shown to be faster than true recall at all but the shortest delays. The recent discovery of the "Moses illusion" (Erikson & Mattson, 1981) is consistent with this view. When people are asked "How many animals of each kind did Moses take on the ark?," most people answer "two" although they know that Noah, not Moses, sailed the ark. This robust finding fits with the idea of a first-stage, fairly automatic selection of potentially relevant facts before the careful second-stage examination. The authors found that the illusion did not hold for Nixon on the same question. Presumably there would be no intersection between an expresident and the story of the ark.

The importance of judging plausibility and reconstruction (which involves plausibility) for memory tasks is part of the explanation of why "passive" reading is so ineffective. The speed with which the search stage can be completed depends on the number of relevant facts. This number hinges on the amount of elaboration that the comprehender has generated. Redundant facts not only allow reconstruction of needed material but they also speed up the inferential processing time for answering questions. Unless the reader elaborates what is read and creates redundant memory structures, not enough information can be retrieved to enable memory reconstruction (Anderson & Reder, 1979).

The comparison of recognition judgments with plausibility judgments was for a restricted domain in these experiments. I believe that the two-stage model shown in Figure 1 can be modified slightly to account for more experimental findings and tasks. If the search stage (for either direct retrieval or plausibility) comes up with no intersection of activation, Stage 2 could be skipped and a fast "no" could be executed. This view is consistent with experiments by Glucksberg & McCloskey (1981). Conversely, if the amount of intersection of activation is extremely large and rapid when plausibility is the strategy selected, then a fast "yes" might be executed inappropriately, as in the case of the Reder and Anderson (1980) thematically related foils.

Reference Notes

1. Chandler, J. P. *STEPIT Program 90PE66.* Bloomington, Ind.: Quantum Chemistry Program Exchange, Indiana University, 1965.
2. Reder, L. M., & Ross, B. H. *The effects of integrated knowledge on question answering: Where and when it helps and where it hurts.* Paper presented at the meeting of the Cognitive Science Society, Berkeley, California, August 1981.
3. Reder, L. M. Unpublished data.
4. Keenan, J. M. Personal communication, September 18, 1981.

References

Anderson, J. R. Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 1974, *5*, 451–474.

Anderson, J. R. *Language, memory, and thought.* Hillsdale, N.J.: Erlbaum, 1976.

Anderson, J. R. Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, 1981, *7*, 326–343.

Anderson, J. R., & Bower, G. H. *Human associative memory.* Washington, D.C.: V. H. Winston, 1973.

Anderson, J. R., & Paulson, R. Representation and retention of verbatim information. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 439–451.

Anderson, J. R., & Reder, L. M. An elaborative processing explanation of depth of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory.* Hillsdale, N.J.: Erlbaum, 1979.

Anderson, N. H., & Hubert, S. Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior*, 1963, *2*, 379–391.

Asch, S. E. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 1946, *41*, 258–290.

Baggett, P. Memory for explicit and implicit information in picture stories. *Journal of Verbal Learning and Verbal Behavior*, 1975, *14*, 538–548.

Bartlett, F. C. *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press, 1932.

Bransford, J. D., Barclay, J. R., & Franks, J. J. Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, 1972, *3*, 193–209.

Bransford, J. D., & Franks, J. J. The abstraction of linguistic ideas. *Cognitive Psychology*, 1971, *2*, 331–350.

Camp, C. J., III. Direct access vs. inferential retrieval across the adult lifespan. Unpublished doctoral dissertation, University of Houston, 1978.

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

Carpenter, P. A., & Just, M. A. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 1975, *82*, 45–73.

Clark, H. H., & Chase, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology*, 1972, *3*, 472–517.

Clark, H. H., & Haviland, S. E. Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension*. Norwood, N.J.: Ablex, 1977, pp. 1–40.

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

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

Dooling, D. J., & Christiaansen, R. E. Episodic and semantic aspects of memory for prose. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 428–436.

Dreben, E. K., Fiske, S. T., & Hastie, R. The independence of evaluative and item information: Impression and recall order effects in behavior-based impression formation. *Journal of Personality and Social Psychology*, 1979, *37*, 1758–1768.

Erickson, T. D., & Mattson, M. E. From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 1981, *20*, 540–551.

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

Graesser, A. C., Robertson, S. P., & Anderson, P. A. Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, 1981, *13*, 1–26.

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

Hayes-Roth, B., & Hayes-Roth, F. The prominence of lexical information in memory representations of meaning. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 119–136.

Healy, A. F., & Kubovy, M. The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition*, 1978, *6*, 544–553.

Johnson, M. K., Bransford, J. D., & Solomon, S. K. Memory for tacit implications of sentences. *Journal of Experimental Psychology*, 1973, *98*, 203–205.

Keenan, J. M., & Baillet, S. D. Memory for personally and socially significant events. *Attention and performance, VIII*. Hillsdale, N.J.: Erlbaum, 1980.

King, D. R. W., & Anderson, J. R. Long-term memory search: An intersecting activation process. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 587–605.

Kintsch, W. *The representation of meaning in memory*. Hillsdale, N.J.: 1974.

Kintsch, W., & Bates, E. Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 150–159.

Lachman, J. L., & Lachman, R. 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 the George A. Talland Memorial Conference*. Hillsdale, N.J.: Erlbaum, 1980.

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

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

Lewis, C. H., & Anderson, J. R. Interference with real world knowledge. *Cognitive Psychology*, 1976, *7*, 311–335.

Mandler, J. M., & Johnson, N. S. Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 1977, *9*, 111–151.

Meyer, B. J. F. *The organization of prose and its effects on memory*. New York: American Elsevier, 1975.

Miller, J., & Greeno, J. G. Goodness-of-fit tests for models of latency and choice. *Journal of Mathematical Psychology*, 1978, *17*, 1–13.

Mohs, R. C., Wescourt, K. T., & Atkinson, R. C. Search processes for associative structures in long-term memory. *Journal of Experimental Psychology: General*, 1975, *104*, 103–121.

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

Pachella, R. G. An interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. Hillsdale, N.J.: Erlbaum, 1974.

Quillian, M. R. Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, Mass.: M.I.T. Press, 1968.

Reder, L. M. The role of elaborations in the processing of prose. Unpublished doctoral dissertation, University of Michigan, 1976.

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

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

Reed, A. V. List length and the time course of recognition in immediate memory. *Memory & Cognition*, 1976, *4*, 16-30.

Rumelhart, D. E. Notes on a schema for stories. In D. G. Bobraw & A. Collins (Eds.), *Representation and understanding*. New York: Academic Press, 1975.

Sachs, J. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 1967, *2*, 437-442.

Schank, R. C., & Abelson, R. P. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, N.J., 1977.

Singer, M. Processes of inference during sentence encoding. *Memory & Cognition*, 1979, *7*, 192-200 (a).

Singer, M. Temporal locus of inference in the comprehension of brief passages: Recognizing and verifying implications about instruments. *Perceptual and Motor Skills*, 1979, *49*, 539-550 (b).

Smith, E. E., Adams, N., & Schorr, D. Fact retrieval and the paradox of interference. *Cognitive Psychology*, 1978, *10*, 438-464.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processing in perception. *Psychological Review*, 1961, *68*, 301-340.

Thorndyke, P. W. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, 1977, *9*, 77-110.

Thorndyke, P. W., & Bower, G. H. Storage and retrieval processes in sentence memory. *Cognitive Psychology*, 1974, *5*, 515-543.

Trabasso, T., Rollins, H., & Shaughnessy, E. Storage and verification stages in processing concepts. *Cognitive Psychology*, 1971, *2*, 239-289.

Received October 23, 1981
Revision received December 16, 1981 ■