Reder, L.M., & Cleeremans, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A. Graesser & G. Bower, (Eds.), *The psychology of Learning and motivation, Volume 25,* New York: Academic Press, pp. 233-258

THE ROLE OF PARTIAL MATCHES IN COMPREHENSION: THE MOSES ILLUSION REVISITED

Lynne M. Reder
Axel Cleeremans

I. Introduction

Despite the fact that there are a large number of models concerned with sentence parsing and comprehension (e.g., Fodor & Frazier, 1978; Frazier, 1987; Just & Carpenter, 1980; Kintsch & van Dijk, 1978) and a large number of models concerned with question answering (e.g., Camp, Lachman, & Lachman, 1980; Graesser & Murachver, 1985; Lehnert, 1977; Norman, 1973; Reder, 1982, 1987; Singer, 1984, 1985), virtually none of these models addresses an important fact about normal sentence or question parsing: People do not pay attention to all of the words that are in the sentence or question. When asked "How many animals of each kind did Moses take on the ark?," most people respond "two" even though they know that it was Noah, not Moses, who took the animals on the ark (Erikson & Mattson, 1981). We misparse sentences like these because often we only partially match sentences or questions to the relevant knowledge structures, failing to note discrepancies. This article is concerned with understanding more about the role of partial matches in comprehension. What factors affect whether we notice discrepancies between the input string and the stored representation of a similar sentence during parsing? Is it easier to ignore discrepancies between the representation of the input and the memory structure or is it easier to find corresponding structures that are very close matches? In this article, we will speculate on

the processes that might be involved in making what we call complete versus partial matches and when and why partial matches are noticed or not noticed.

The Moses Illusion, as it is called, was first demonstrated by Erikson and Mattson (1981). They found that people frequently failed to notice a distortion when asked to answer a question or verify an assertion such as *Moses took two animals of each kind on the ark*. The authors were careful to confirm that people who fell for the illusion did in fact know that it was Noah, not Moses, who took the animals on the ark. Even when people were warned that there might be some "tricky" or distorted questions, there was still a large tendency not to notice the distortions until they were pointed out.

A common reaction to this result is that people are just cooperating with the speaker. That is, there is the view (Grice, 1975) that people know what is meant by a question and therefore simply behave in a way that reflects the shared knowledge. We contend that the behavior of subjects is not caused by a conscious decision to be "cooperative." On the contrary, people who *notice* the mismatch would probably respond "you mean Noah" rather than just respond "two." One of the goals of this article is to examine the conditions under which people notice distortions or fail to do so.

The experiments in this article provide further evidence that the typical finding, that people ignore the distorted portion of a question, does not reflect a tendency to "cooperate" on the part of the listener. Rather, it reflects a natural tendency to fail to notice the mismatch. These experiments ask subjects to ignore distortions in questions and answer them as if the questions were not distorted. We compare this task with the situation where subjects are asked not to answer the question when it is seen in distorted form. In other words, the experiments compare the ease of making careful versus partial matches to memory.

In order to better understand when and why people use partial matching to memory structures, this article also focuses on some of the variables that may affect a person's tendency to use partial matching. For example, it is possible that level of activation of the relevant memory structures may play a crucial role in the completeness of the inspection of that structure. Does the tendency to fall for the Moses Illusion depend on the number of terms in the query that match the memory structure, or does it depend primarily on the similarity of the original to the substituted term? Does familiarity with the relevant knowledge structures play a critical role in a person's susceptibility to the illusion? Are we more or less likely to ignore the mismatch because the information is highly familiar? Do these variables interact, or affect separate stages in processing?

TABLE I

EXAMPLES OF FOILS USED IN EXPERIMENTS

What kind of tree did Lincoln chop down?
What did Goldilocks eat at the Three Little Pigs' house?
Who said, "Ask not what you can do for your country, but what your country can do for you."?
What phrase follows "To be or not to be" in Macbeth's famous soliloquy?
What month is associated with Mother's Day, Veteran's Day, and spring flowers?
By having an apple fall on his head, what did Galileo discover?

In the experiments by Erikson and Mattson (1981) only a few questions were asked of subjects, most of which were distorted. In the experiments to be reported here, subjects were asked to answer a large number of questions, only half of which were distorted. Examples of distorted forms of these questions are presented in Table I. In order to compare the relative ease of partial versus complete matches, subjects were sometimes instructed to ignore slight distortions in questions (e.g., substitution of Moses for Noah), while at other times they were instructed to say "can't say" if the query contained a substitute, i.e., a related, but inappropriate, term that made the question nonsensical or unanswerable if treated literally. Subjects were instructed that half of the questions would be distorted and half would not be distorted. When subjects were in the *literal* task, they were to treat each question literally and not give an answer if the question had been distorted; however, when subjects were given *gist* instructions, they were to ignore minor distortions and give answers to questions as if they were not distorted. The various experiments differed in some respects, but all used these materials and had these basic task characteristics. As will become apparent, the general results are quite robust. As a result, not all replications of this phenomenon will be reported.[1]

There are some basic predictions about the ease of this task that derive from assumptions about the relative difficulty of partial versus complete matches: If complete matches between a test probe and a memory struc-

[1] For example, an experiment was conducted comparing young college alumni with retired college alumni. The same basic pattern was found for both groups. The only difference between the age groups was that older subjects showed bigger effects. Likewise, Experiment 2 is very similar to another study we conducted that is not reported for space reasons. Therefore, all statistical analyses should be considered underestimates of the reliability of these results.

ture are easier to compute than partial matches, then it should be easier to say "can't say" to a distorted question than to give the answer to it. On the other hand, if normally we only partially match probes to memory structures, then when we are forced to make careful inspections of memory, the process should be costly, making the literal task harder than the gist task.

II. Experiment 1: The Moses Illusion When There Are Many Tricky Questions

A. Method

1. *Subjects*

Some of the subjects were Carnegie Mellon alumni who had finished their undergraduate education less than 5 years prior to participating in the experiment. These subjects were run as a control group for an experiment that looked at the effects of age on various tasks. The other subjects were enrolled as undergraduates at Carnegie Mellon and participated in order to help fulfill a course requirement.[2] Both groups had to answer questions, half of which were distorted, as illustrated in Table I. There were 18 alumni and 16 students in the literal condition, and 19 alumni and 14 students in the gist condition.

2. *Design and Procedure*

Subjects were randomly assigned to either the literal condition or the gist condition. The former group was asked to discriminate between distorted questions, i.e., questions where one of the terms had been replaced with a related but inappropriate term, answering "can't say" to distorted questions and giving the correct answer only to undistorted questions. The latter group was asked to ignore distortions in questions and answer either version of a question as if there were no substitution of terms.

Two versions were made of each question, and subjects saw only one version of a given question. The assignment of questions to distorted versus normal was made randomly for each subject with the constraint that half of the questions be distorted and half normal. There was one additional factor in the experiment: We varied the number of terms in the statement that were associated with the answer, either few or many associated terms. This variable will be discussed further in the section on materials.

[2] The latter group participated in a within-subject design such that they received the literal task instructions for one block of trials and the gist instructions for the other block. For purposes of this article, we will only consider their performance in the first block and thereby treat the experiment as a between-subject design, viz., literal group versus gist group. The excluded data (from the second trial block) was essentially indistinguishable from that from the first trial block. We decided that we would prefer to include more subjects (viz., young alumni and college students) than treat the experiment as a within-subject design.

A subject was instructed about the nature of the task according to the condition he or she was assigned to. Both groups of subjects were told that they would see one question at a time on the computer screen and be asked to give their answers through the microphone as quickly as possible while maintaining accuracy. In the gist condition, subjects were told:

> Some of the questions are improperly constructed and, if taken literally, do not have an answer; however, we want you to get the gist of the question and give the best answer you can.

Subjects in the literal condition were told:

> Some of the questions are improperly constructed and, if taken literally, do not have a correct answer. Please treat each question literally and answer "can't say" to the ill-formed questions.

Subjects controlled the rate at which each new question was presented by pressing a NEXT button on a button box. On each trial, a question would appear on the screen, which started a clock. A subject's verbal response into the microphone would cause the clock to stop and the question to be erased from the screen. The program would automatically record the time for that trial and the experimenter would type into the computer the response the subject gave. Both groups were instructed to say "don't know" if they did not know the answer. Subjects received feedback in that the computer displayed the expected answer on the screen. The experimenter was only required to type in the subject's answer if it differed from the one displayed. For example, if the question was *What kind of tree did Lincoln chop down?*, the answer *cherry* would be displayed on the screen for the gist group after a response, while the answer *can't say* would be displayed on the screen for the literal group. Of course, if the question had used *Washington* rather than *Lincoln*, both groups would see *cherry* as the answer.

The experimenter was present at all times, both to type in the subject's response and to ensure that the voice key was not triggered accidentally by an unintended vocalization. The experimenter also noted if the re-

sponse time was invalid due to a subject speaking too softly to trigger the voice key.

3. *Materials*

Sixty-two pairs of questions, like the Noah–Moses question, were constructed using the following constraints: (1) the term to be substituted in the question had to be semantically confusible with the original term; (2) it had to come from the same part of speech (syntactically); (3) it had to share some phonetic features with the term, and (4) the distorted question should not be interpretable in a different manner such that there would exist a different correct answer. For example, we could not use the following distorted question: *On what holiday do children go door to door, dressed in costumes, saying "Ho, ho, ho"?* The intended answer is *Halloween*, but the expression "ho, ho, ho" is associated with Christmas, and therefore that question would produce a competing response to the intended answer *Halloween*. Therefore, the distorted version of the Halloween question was written in the following form: *On what holiday do children go door to door, dressed in costume, giving out candy?*

It was also essential that the base form of the question be answerable in the absence of either the original or the substituted term. For example, *toll booth* was substituted for *phone booth* in the question *What comic book hero does Clark Kent become when he changes in a toll booth?* The substituted term invalidates the question but is irrelevant to figuring out the answer if the term is ignored. The substituted term did not always appear at the end of the sentence. Whether a distorted version of a question could create the illusion depends on the listener's knowledge about the topic and the two terms being substituted. Knowing too much or too little makes it very difficult to create the illusion. This issue will be addressed later in the article.

The pairs of sentences, one properly formed and one distorted, were essentially identical in length; however, the set of sentences used varied in length and, more importantly, in the number of content words associated with the answer. Some sentences were rated as having only two words that are semantically related with the answer, and some were rated as having as many as six words associated with the answer. Three independent raters judged these sentences for the number of content words, and there was almost no disagreement among them. Half of the sentences had two or three related terms and the other half had four, five, or six related terms. The former half we call the *few-terms condition*, and the latter we call the *many-terms condition*.

A pilot experiment was run on 15 subjects to screen our materials. We had to eliminate and replace a number of questions for various reasons.[3] Later in the article, we will discuss the factors that make a question a candidate for the Moses illusion.

B. Results and Discussion

The response-time data are displayed in Fig. 1A for correct responses that do not involve an inaccurate measurement. Approximately 4% of the trials were discarded because the subject either stopped the clock early (e.g., by coughing into the microphone) or because the microphone did not pick up the articulation. An analysis of variance was done on the Subject Type (alumnus vs. college student) × Answer Condition (gist vs. literal) × Number of Related Terms (few vs. many) × Question Type (normal vs. distorted) for response time and accuracy.

The data are collapsed over the two types of subjects, alumni or current undergraduates, since there were virtually no effects due to subjects and this variable was not of theoretical interest. Figures 1A and 1B present the response-time and accuracy data, respectively, as a function of number of terms, task, and distortion. First consider the response times (RT). There is a sizeable and highly significant RT advantage for the gist condition compared with the literal condition, $F = 6.14$. There was no significant difference in RT for distorted vs. normal sentences, $F < 1.0$. The number of words in the question that were associated with the answer also affected response time, $F = 69$, such that subjects were significantly slower when there were more associated terms. By itself, this result would not be interesting since this variable is confounded with sentence length; however, the accuracy data also show an effect of this variable.

Almost all manipulations had an effect on the accuracy measure. First, subjects made significantly more errors when asked to parse questions literally than when asked to give a gist response, $F = 65$. This effect is due primarily to subjects in the condition where they are required to say "can't say," namely, the distorted literal condition. Subjects made significantly more errors in that condition than any other, resulting in a significant, ($F = 61$), Question Type × Task interaction. The fact that the error rate is so much larger for distorted questions in the literal condition means that the

[3] Some were rejected because not enough people knew the answer, e.g., *What is the name of the woman who is opposed to ERA, adoption, and other liberal causes?* (Answer: *Phyllis Schlafly*, substitute *abortion* for *adoption*.) Others had to be rejected because subjects would give the wrong answer in either form of the question, e.g., *What corner of the envelope should one put the return (zip) code?* The answer is supposed to be *upper left;* however, subjects would always respond "upper right," assuming that the question concerned where one places the stamp.

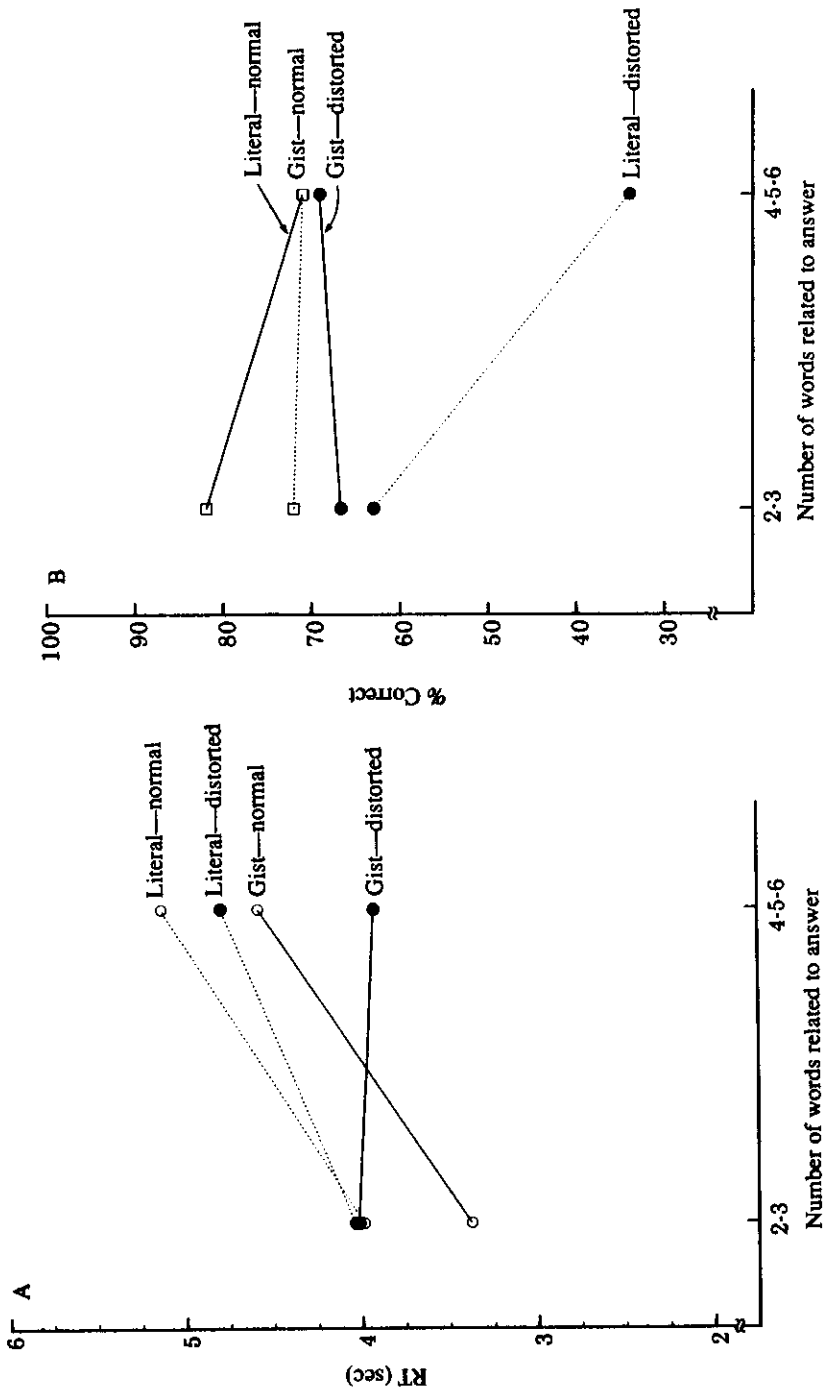

Fig. 1. A, mean correct response times (sec) and B, mean accuracy (%) as a function of the task (gist or literal), the number of words in the question associated with the answer, and whether the question was distorted or undistorted (normal), in Experiment 1.

difference in response times between the literal and gist tasks is an underestimate. That is, if subjects tried to be more careful in the literal condition (so that accuracy would be as good for the distorted questions as for undistorted questions), responses would be slower still in the literal condition.

Finally, the triple interaction of Question Type × Task × Number of Related Terms was also significant for accuracy, $F = 13$, such that the number of words in the question that are associated with the memory structure had no impact on performance except in the literal task when the question was distorted. Subjects generally were least accurate with distorted questions in the literal task (as compared with undistorted questions or the gist task), but the situation was exacerbated by having many terms in the question associated with the relevant memory structure. It may just be that it is harder to check every term in longer sentences, but subjects seemed to have a bias to give the undistorted answer since their performance was unaffected by many terms if the question was undistorted.

One conclusion that seems clear from these data is that it is easier to make partial matches than complete matches. This pattern is very robust: We have replicated it in five similar versions of this basic experiment, not to mention the original Erikson and Mattson (1981) result. On the other hand, it is less clear what affects the acceptability of a partial match. That is, not every distorted sentence will be incorrectly accepted. The acceptability of a partial match is affected not only by the number of terms that match the memory structure (as demonstrated in this experiment), but it is also affected by the similarity of the substituted term for the correct one (choice of the distortion term). Erikson and Mattson found, for example, that subjects never false alarmed to the question *How many animals of each kind did Nixon take on the ark?*, even though *Nixon* shares the same number of syllables and same first phoneme with *Noah*.

Consider the following statements taken from Bredart and Modolo's (1988) work on the Moses Illusion. These are translations of the French statements that their Belgian subjects had to verify or reject:

> *It was Cinderella who was sheltered by seven dwarfs before marrying her prince.*
> *It was Magellan who discovered America at the end of the 15th century.*
> *It was the Jewish physicist Max Planck, author of the theory of relativity, who emigrated to the United States.*

We seriously doubt whether any American subjects would false alarm to the statement that substitutes *Magellan* for *Columbus*. Probably the

strength of the assertions being tested affects the tendency to find the Moses Illusion. American and Belgian subjects differ in their respective knowledge concerning Columbus and Magellan. Most Americans know more about Columbus, and probably less about Magellan. Conceivably for Belgians, the principal features associated with both Columbus and Magellan are that they were ocean explorers of the New World centuries ago. Similar arguments can be made for the Einstein–Planck question.

This difference in susceptibility to the illusion based on prior knowledge suggests that if too much is known about a term that is to be replaced, the illusion will not work. Informal tests of the Moses Illusion with devout friends who attend Bible classes were consistent with this view: People who know the Bible are much less likely to confuse Moses with Noah.

It occurred to us that if our subjects were highly familiar with the information that they were to be tested on, then they would find it easier to discriminate distorted from nondistorted sentences. For example, if someone were to be asked a question such as *How many children did your mother, Anna, have?*, the answerer would notice that Anna is not the mother's name even if the mother's name were similar, e.g., Ann. The next experiment investigates this possibility.

III. Experiment 2: The Moses Illusion with Highly Familiar Facts

In this experiment, we asked subjects to commit to memory a series of facts prior to answering questions. These facts were a subset of those facts necessary to answer questions in the "Moses task." After the subjects had committed the facts to memory, the remainder of the experiment was identical to the previous experiment (Experiment 1) in terms of the procedure. The difference was that half of the questions had recently been primed by having the subject either read or memorize the relevant information. Thus we might prime a question by having the subject commit to memory the correct form of the fact, e.g., *Noah took two animals of each kind on the ark*. Regardless of whether the question would be assigned to the distorted or normal condition, the sentence to be studied was always of the correct form. So a subject might later be asked "Who does Clark Kent become when he changes in a toll booth?" but the studied sentence was *Clark Kent becomes Superman when he changes in a phone booth.*

A. Method

The materials, in terms of the questions, were exactly the same as in the previously described experiments. We constructed priming statements

based on the questions that simply combined the answer and the undistorted form of a question into a declarative statement. We attempted to make the form of the statement very similar to clauses within the question. Half of all the questions were randomly selected to be primed for each subject, with the constraint that half of the primed questions be from the set selected to be distorted and half be from the remainder, i.e., those left undistorted. Subjects always studied the correct or undistorted statement during the priming phase. This means that the match between studied statement and question was not as close in the condition involving questions that were both primed and distorted.

Subjects read through the priming statements carefully. They were told to try to memorize them, but they were not told that the statements they were studying would be involved in a later experiment. Subjects not only had to carefully study the statements, they had to be able to recall them perfectly to a cue word from the sentence. This was done in a "drop-out" procedure. This procedure involved presenting the cue or sentence topic and asking the subject to recall the studied sentence verbatim. If the sentence recall was perfect, that sentence dropped out of further study trials. Otherwise the statement was represented for study. This was followed by a test (cue-word prompt) from one of the other not-yet-recalled statements. This cycle would continue until all statements were recalled.

The procedure during the Moses Illusion questioning phase was very similar to that of the previous study. The experiment was a within-subject design where subjects were randomly assigned to either get the literal task first or the gist task first. There were 30 subjects; 14 performed the gist task first, and 16 performed the literal task first. This order variable had no impact on the data and we therefore collapsed over it in all analyses.[4]

B. Results and Discussion

The response time and accuracy results of the "memorize" experiment are displayed in Figs. 2A and 2B as a function of whether or not the information was primed, whether it was distorted, and whether the task was to treat questions literally or in a gist fashion. Accuracy is again defined as those correct of those attempted, i.e., we do not count trials where subjects felt they did not know the answer or the voice key did not work.[5] An arcsine transformation was done on the accuracy data before

[4] We also conducted another experiment where subjects studied the statements but did not commit them to memory. The results were very similar although somewhat less dramatic. We do not report that experiment for reasons of space.

[5] We will later present an analysis of how many trials were "don't know" trials as a function of condition.

submitting it to an analysis of variance. Consider the response-time data first, displayed in Fig. 2A. The basic pattern matches what we found before, namely that the literal task is slower than the gist task, $F = 35$. The distorted questions take slightly longer to answer than the undistorted, $F = 10$, but there is no interaction of task and distortion.

Of more interest is the effect of priming on these answer times. It is not surprising that there is a significant effect of priming such that previously studied statements are answered much faster than those that were not studied, $F = 49$. This may be due to an encoding advantage of having recently seen and parsed the statement. Subjects were also more accurate for studied statements, $F = 123$. Having just memorized the critical fact, subjects obviously knew the answer to that question.

More surprising is the fact that the familiarization variable appears to be additive with other variables. For response time, there was virtually no interaction between making the information more accessible and the ease or difficulty of doing a careful match versus a partial match. Priming resulted in a full 1-sec savings in response time, regardless of whether or not the question was distorted. It seems that priming the statements did not make it easier to detect the mismatch. If priming had affected the matching process, then the improvement in error rate due to this variable should have been greater for the literal distorted condition than for the literal undistorted condition. The same argument could be made for the change in response times. Finally, if familiarity impacted on the match process, the gist condition would be expected to suffer since distortion should be more salient when the knowledge has been primed. That means that priming would make it easier to notice distortions and harder to ignore them.

The additive effect of familiarization on this task can also be seen in the accuracy data. The intersecting lines in Fig. 2B reflect the fact that both the gist and literal tasks perform well with normal questions, but the literal task suffers much more than the gist task with distorted questions. This is true regardless of whether or not the question was previously studied. In other words, there appears to be virtually no interaction between making the information more accessible and the ease/difficulty of doing a careful match versus a partial match.

This conclusion is supported by another analysis. We looked at the types of errors that were made in the various conditions in order to distinguish between errors due to lack of knowledge and errors due to incomplete matching between the input and the memory structure. There are four types of errors when performing the literal task: a subject can (1) say "don't know," (2) give the undistorted answer when the question was distorted, (3) say "can't say" when the question was not distorted,

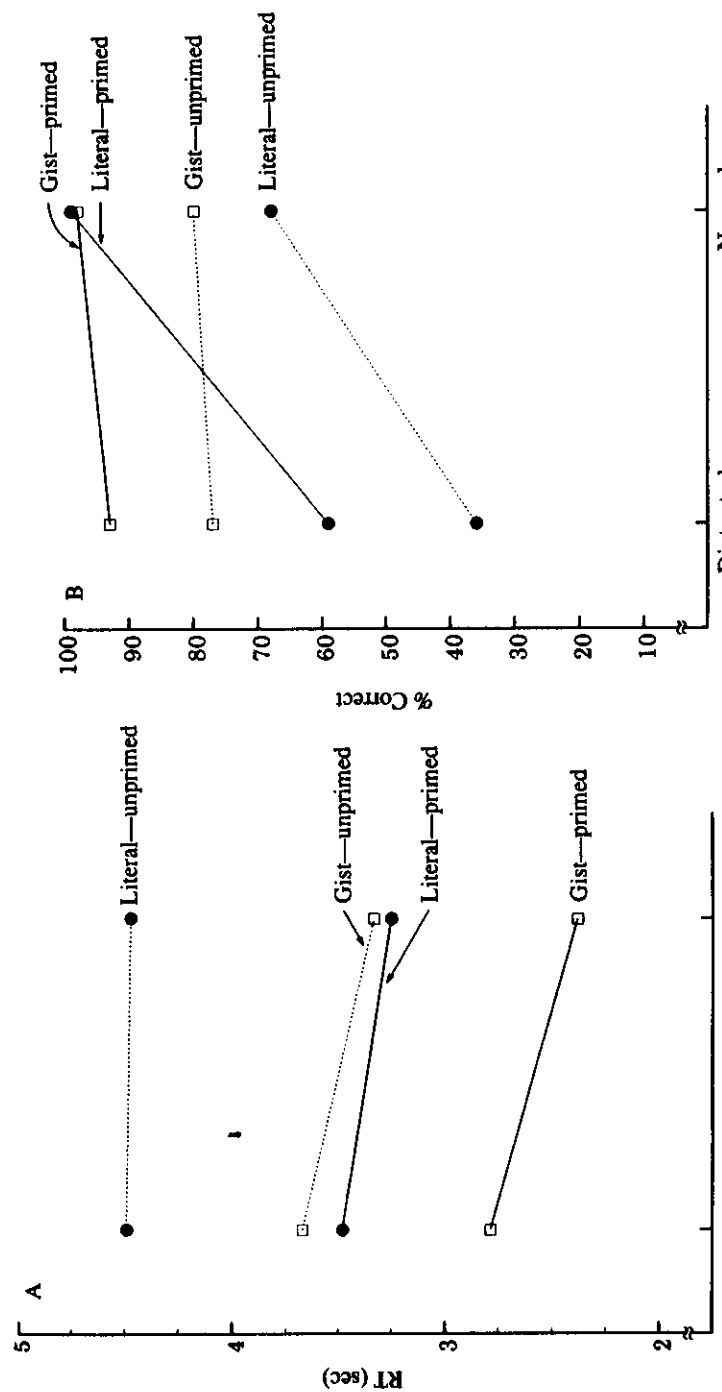

Fig. 2. A, mean correct response times (sec) and B, mean accuracy (%) as a function of the task (gist or literal), whether or not the answer was a primed, and whether the question was distorted or undistorted (normal), in Experiment 2.

or (4) give an unrelated, wrong answer. In the gist condition, errors fall into only two categories, "don't know" and "wrong answer."

First consider this analysis for the experiments that did not involve priming. The error data from Experiment 1 are displayed in Table II as a function of type of error. Notice that the proportion of errors for "wrong answer" and "don't know" is essentially the same for both the literal task and the gist task regardless of whether or not the statement was distorted. This confirms one's intuition that the larger error rate in the literal task (compared with the gist task) is due to giving the undistorted answer to distorted questions and to occasionally saying "can't say" to a question that was not distorted.

Now consider the experiments where half of the test questions had been primed. Table III presents the same analysis except that the data are also partitioned into the primed and unprimed categories. For the unprimed data, the numbers look very similar to those in Table II. Subjects do not know the answer to about 20% of the queries. These data are in stark contrast with those from the primed conditions. In the primed conditions, subjects only say "don't know" to about 3% of the questions. That shift is of course reasonable since they have just been studying the answers. On the other hand, subjects give nearly as many undistorted answers to distorted questions in the familiar condition as in the unfamiliarized conditions.

To summarize the results from this experiment, we found that the basic Moses Illusion was not affected by memorizing the information relevant for answering the questions. The manipulation of studying half of the facts prior to answering the questions had a large impact on error rate, but

TABLE II

ERROR RATES, BY TYPE, FROM EXPERIMENT 1[a]

	Literal task		Gist task	
Type of error	Normal	Distorted	Normal	Distorted
Wrong answer	.07	.04	.03	.10
Don't know	.14	.13	.15	.20
"Can't say"	.11	n.a.	—	—
Undistorted answer	n.a.[b]	.33	—	—
Total	.32	.50	.23	.30

[a] These data are only from the college students. For technical reasons it was impossible to reanalyze the data from the alumni.

[b] n.a., not applicable.

TABLE III

ERROR RATES, BY TYPE, FROM EXPERIMENT 2

	Literal task				Gist task			
	Normal		Distorted		Normal		Distorted	
Type of error	Primed	Unprimed	Primed	Unprimed	Primed	Unprimed	Primed	Unprimed
Wrong answer	.004	.06	.01	.04	.02	.06	.06	.09
Don't know	0	.18	.01	.20	0	.14	.01	.15
"Can't say"	.01	.08	n.a.	n.a.	—	—	—	—
Undistorted answer	n.a.[a]	n.a.	.34	.40	—	—	—	—
Total	.01	.32	.36	.64	.02	.20	.07	.24

[a] n.a., not applicable.

mostly on accessibility of the relevant knowledge: not surprisingly, subjects were faster and more accurate on the studied information. On the other hand, the probability in the literal task of giving the undistorted answer to a distorted question was essentially unaffected by whether the information had been primed and made more accessible.

IV. General Discussion

The first experiment replicated the basic results of Erikson and Mattson (1981) in a task where subjects knew to expect distorted questions. It was found that subjects find it very difficult to detect and report distortions in the questions that they must answer. This difficulty of noting distortions in the memory trace is especially difficult when there are many terms in the probe that do match the memory trace.

It appears that the processes that are affected by the task variable and the distortion variable are distinct from the processes that are affected by studying the relevant facts. In other words, it seems that the effect of priming on accuracy occurs exclusively in a reduced tendency to say "don't know" or give a wrong answer. There was almost no effect of priming on the tendency to give the undistorted answer in the distorted literal condition. In contrast, the task variable (gist vs. literal) did not affect the tendency to say "don't know" at all even though this variable had a large impact on the error rate.

In other words, the manipulation of priming seems to affect the accessibility of the knowledge, both in terms of speed of access and whether the

fact is accessed at all, but it apparently does not change the ease of making careful matches.

What do these results say about our normal, everyday comprehension of text and other forms of input? We think it is clear that the default mode of processing is as effortless as possible. People try to do as little work as possible to comprehend or understand. It is clear that much of our listening is expectation based. These expectations are formed at all levels: We tend not to notice distortions at the phonemic level (e.g., MacWhinney, Pleh, & Bates, 1985; Warren & Warren, 1970), at the syllable or word level, or even at the phrase level. For example, few people would notice the distortion of the expression *Ask not what you can do for your country, but what your country can do for you*. It is less clear whether we rely on our expectations because it is so difficult to encode and check the input with stored information, or whether we rely on our expectations simply because it is easy and usually works.

How did processing change in our experiment as compared with "normal" processing? In nonexperimental situations, a person will typically set a loose criterion for finding a relevant memory structure to match an input. In the literal condition of these experiments, subjects readjusted their criterion upward so as not to be easily tricked. This readjustment was caused both by the explicit instructions to be careful and subjects' experiences of failure to notice distortions and consequent errors. Even with these reminders, subjects continued to make far more errors in the literal condition and to respond more slowly than in the gist condition.

What does this criterion refer to? One possibility is that this criterion reflects the number of tests to accept a match in memory. When it is very important to be careful, a number of additional tests are probably made between an input probe and a memory structure. For example, all the relational information in a memory structure might be tested. There is evidence (e.g., Ratcliff and McKoon, 1989) that the relational information is available at a later time than the simple match of features. If the relational information is not tested, the question *On which holiday do children, dressed in costume, go door to door, giving out candy?* would easily slip by as acceptable. Other tests would involve making sure that each word in the query matched the memory structure that contains the answer.

Any model that is to account for these data must include an adjustable criterion for acceptability of a match. Note, however, that this adjustable criterion does not refer to a threshold for finding a candidate memory structure. This is because the data from the priming studies suggest that studying the relevant memory structures affects the ease of access of the information but it does not affect the ease of making complete matches to memory structures. In other words, learning more about the information

and/or having it be more accessible does not have an impact on one's ability to notice distortions and therefore the ease of saying "can't say."

When the criterion for a match is set low, it still seems that the standard for acceptability of a match is not a simple proportion of concepts that match between the probe and the memory structure. Rather, certain aspects of the probe and memory trace are more critical to match, while others are seen as details that are less critical. The given/new distinction, a construct from discourse processing, should be a dimension that affects what is critical to match: Given information is assumed to match without careful checking. People are more inclined to carefully match those aspects that are thought of as new.

The Moses Illusion has been found in other forms, for example, people have the illusion of knowing (e.g., Glenberg & Epstein, 1985, 1987; Glenberg, Sarocki, Epstein, & Morris, 1987) or understanding when, in fact, the passage they are reading is replete with contradictions. The illusion of comprehension is greater when the violation or contradiction is contained within the given rather than the new information. People tend to focus on the new information while the assumption is that the given information is given and therefore is correct. This probably also explains why the Moses Illusion is much larger in question form than in verification form: A person focuses even more on the targeted information when asked to answer a question pertaining to the information queried.

The research by Bredart and Modolo (1988), briefly described earlier, that replicated the Moses Illusion with Belgian subjects, also looked at the effect of focus on the size of the illusion. They contrasted statements where the distorted form either was the focus of the sentence or was not the focus, e.g., *It was Moses who took two animals of each kind on the ark* vs. *It was two animals of each kind that Moses took on the ark*. Not surprisingly, the illusion was greater when the distortion was not in focus.

Erikson and Mattson (1981) claimed not to find effects of focus; however, their subjects were not expecting "tricks" and they had not been sensitized to raise the criterion for goodness of matches. On the other hand, their definition of focus was simply whether it was in question form (unfocused) versus statement form (focused). Even in their case, however, the effects were greater (but not significantly so) for the unfocused or question form.

We are still left with a few questions. Consider again the question *In what year did Magellan discover America?* Although we have no data to back up our claims, we are confident that most Americans would not false alarm to this question. Likewise, we believe that biblically trained people will not false alarm to the Moses question. Yet, training subjects to learn the relevant information did not affect their tendency to accept distorted

we would train the network by presenting it with a set of undistorted statements like *Noah took two animals of each kind on the ark*, and its task would be to reproduce them on its output pool. After training, we can then test the network by presenting it with incomplete patterns and measure how well it is able to reproduce the corresponding complete statements. Thus, we expect the network to produce a representation of the whole statement when presented only with the concepts corresponding to the question form of this statement. In other words, the network should perform pattern completion and produce *Noah took two animals of each kind on the ark* when presented only with *How many animals did Noah take on the ark?* At this point, we have shown how the network can be trained to answer *undistorted* questions. Understanding how the network performs with *distorted* questions requires examining the assumptions underlying the representation of input information in more detail.

Assume that the input information is represented by a large number of microfeatures, that is, each concept in a question would be represented by a pattern of activation on some subset of the input units. Under the condition that such distributed representations (Hinton, McClelland, & Rumelhart, 1986) are used for the input information, similar concepts (like *Moses* and *Noah*) can be represented by overlapping patterns of activation. Because each unit in the network contributes to the production of the output, similar input patterns will tend to result in the emergence of similar outputs. Thus, to the extent that the overlap between the representations of a concept and its distorted version is large enough, the answer to the question will still be available (although in a somewhat weakened fashion) when the network is presented with the distorted version of the question. This type of model naturally handles the basic Moses Illusion because it gives the answer as though the question were not distorted.

To account for the basic results in the gist task, let us further assume that reaction time and accuracy are proportional to the error associated with the output.[7] After training, each response of the network will be associated with a specific error. It will then be possible to find a criterion in the error under which it is assumed that retrieval is successful. Large distortions would entail the error to rise above that criterion and the output to be garbled (leading to "don't know" or wrong responses). Small distortions would only weaken the output, thus leading to slightly slower response

[7] For instance, we could assume that a response is chosen with a probability and a latency proportional to its strength. Strength could be defined in terms of response competition and could be evaluated by the Luce ratio of the activation of the relevant units. Obviously, specifying a complete model of how reaction times and accuracy are related to the output of the network is outside the scope of this discussion.

times, as was indeed observed. We can thus account for the basic fact that in the gist condition subjects are able to give the answer to distorted questions as long as the distorted and normal concepts are similar enough, as measured by the number of overlapping features between their representations. Figure 3 illustrates this point by showing how the network would respond when presented with the question *How many animals did Moses take on the ark?*

Other experimental results can also be understood by assuming that memory processing is essentially similarity based. Three of these results can be approximated within the PDP framework: the effect of the number of related terms, the effect of focus, and the effect of prior knowledge of the domain.

1. *Number of Related Terms*

It was shown that subjects find it increasingly difficult to detect distortions when the number of concepts that are associated with the answer increases. In this model, the effect can be explained by the ratio of matching to mismatching features. The output will be more extensively disrupted when the ratio of matching to mismatching features is small than when it is large. In other words, with fewer related terms, the overall proportion of distorted priming features is larger, making it harder to retrieve the answer. Thus, the model accounts for the fact that the illusion is harder to detect when the number of priming concepts in a question is large.

2. *Focus*

The focus effect refers to the fact that the illusion is less likely when the distorted term is the focus of a sentence. In terms of a PDP model, the effect can be approached by giving more processing importance to the focused concept. This could be achieved in a variety of ways. For instance, one could train the network so as to bias its processing toward the pool of units holding the representation of the concept in focus (Cohen, Dunbar, & McClelland, in press, do something similar). Concepts presented under focus will then have more impact on the output than the other concepts because the activation of the "focus" units will bias the impact of the activation of the input units in processing. Distorting the concept under focus will therefore entail a larger degradation of the output than if the distortion is limited to a concept not under focus, thus making it easier to detect the illusion in the former case. Similarly, both the fact that subjects focus on new rather than on given information and the fact that the Moses Illusion is greater in question answering than in sentence verification can be explained by the presence of linguistic markers that redirect subjects'

questions as undistorted. Learning the information only affected accessibility. What then is the difference? We suspect that for experts in a field (e.g., Americans with the Magellan example and biblical scholars with the Moses example), it is not the fact per se that is so much better known, but that the differentiation of features between the distorted and undistorted term is much more articulate. For us, Columbus is a very rich concept and we can easily distinguish it from Magellan. A biblical scholar can easily see the differences between Noah and Moses, while to most of us they share the features of "ancient, Bible-story characters."

A. PDP Systems: A Potential Framework for Modeling the Moses Effect

The above discussion raises a number of important issues regarding the processing features of memory. It seems clear that any model of memory processing must possess a number of critical properties such as (1) expectation-based processing (i.e., automatic generation of default values and prototyping), and (2) graceful degradation of performance in the presence of incomplete or distorted information. In this section, we describe one approach for modeling the various effects we have examined.

There are several reasons to choose the class of architectures called parallel distributed processing (PDP) systems as the framework for attempting to model these effects. The PDP framework has been widely applied to simulating memory phenomena in general (e.g., McClelland & Rumelhart, 1986), and its similarity-based processing features seem a priori very relevant for the Moses effect. Indeed, one of the most appealing features of PDP models is the fact that their processing is fault tolerant, that is, the system always attempts to come up with its best guess about the identity of the probed information. Thus, incomplete information and noisy cues tend not to disrupt memory retrieval as long as the distortions are small enough.

How does the PDP framework model the basic Moses Illusion? Let us consider, for the sake of clarity, a simple hypothetical network consisting of three interconnected pools of units. The first pool of units is used to represent input information. All the units of this pool are connected to all the units of a second pool, and all the units in this second pool are connected to all the units of a third, output, pool that holds representations of the output of the network. Processing in such a system consists of presenting the network with a pattern of activation representing the input information and letting the activation spread through the connections up to the output layer. The desired mapping between input and output is achieved by *training* the network. Figure 3 shows the general architecture

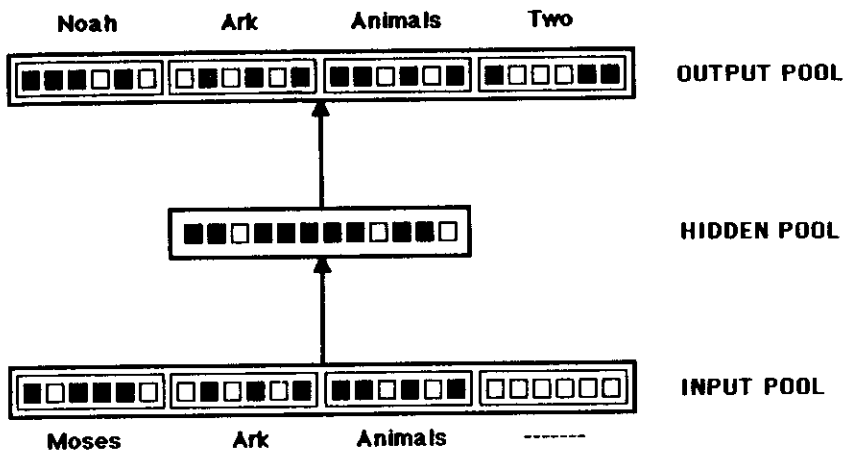

Fig. 3. A simplified representation of a hypothetical pattern-completion network. The network is shown processing a distorted question at test time. The input pool holds distributed representations of the concepts of this distorted question (*How many animals of each kind did Moses take on the ark?*). Each concept is represented by a pattern of binary activation values on a subset of the input units. The units of the output pool represent the responses of the network. In this example, this pool holds a weakened version of the undistorted statement *Noah took two animals of each kind on the ark.*

that we used (in this case, the network is shown processing a distorted question at test time, i.e., once training is completed). Training consists of repeatedly presenting the network with the input/output pairs that need to be learned. On each presentation, the connections are modified in such a way as to reduce the difference (or error) between the actual and target output.[6] The "error" is classically defined as the sum of the squared differences between the target and actual activation of the output units. Training stops once the total error (i.e., over all the input/output patterns) drops below a specified threshold. Once training is achieved, the network will have developed internal representations (on the second, "hidden" pool of units) that allow it to produce the desired output when presented with the corresponding input pattern. In order to model the Moses effect in such an architecture, we first need to give the network knowledge about a set of undistorted statements. A simple and elegant way of doing so is to assign the network the task of reproducing on the output pool the same information that it is presented with on the input pool. In the present case,

[6] Using, for instance, the back-propagation learning algorithm (see Rumelhart, Hinton, & Williams, 1986).

attention. In terms of the PDP framework, both of these effects might be approached in the same way as focus itself.

3. *Knowledge of the Domain*

The fact that experts in a given domain are not easily tricked by questions pertaining to that domain is probably due to the fact that expert representations are richer than those of novices. Concepts will tend to be represented by more features in the expert's mind and, presumably, these additional features tend to discriminate between the relevant concepts. In other words, the proportion of shared to overall number of features between two concepts like *Moses* and *Noah* will tend to be smaller for experts than for novices. As a result, and in terms of the PDP framework, these more differentiated representations will tend to have contrasting effects on the output and to facilitate the detection of the illusion.

B. Challenges to a PDP Approach: The Literal Task and the Distinction between Retrieval and Matching

Up to this point, we have only considered a situation similar to the gist condition in the reported experiments, in which the system only has to report the answer to the question, whether the question is distorted or not. This discussion suggested that a PDP network using distributed representations, and trained to perform pattern completion, could be used to produce the answers to a set of questions. We also showed that its responses would remain readable even if some parts of the representations associated with the questions are distorted, thus reproducing the basic Moses Illusion in a natural way.

However, the task facing subjects in the literal condition is very different. Indeed, the task requires distortions to be detected in the course of processing in such a way that the system can report them ("can't say" responses). Another important feature of the literal task is the increased number of possible responses. In the gist condition, there are only two classes of possible responses ("don't know" and "wrong answer" versus "undistorted answer"). The literal condition allows three different types of responses (those from the gist condition, plus "can't say"). In terms of our model, these three (possible) types of responses must be distinguishable from each other. Finally, our priming results indicate that detecting mismatches and retrieving information are distinct processes.

How can this task be simulated within the PDP framework? The answer to this question is far from obvious. The problem stems from the fact that we need to be able to distinguish between failure to retrieve and mismatch (i.e., between "don't know" and "can't say" responses). The distinction

should be based on some simple measure of the network's performance. Assume that "don't know" answers occur when the network simply fails to produce any interpretable pattern on the relevant output units when presented with a given question (most of the output units could have near-zero activation, for instance). This type of response can easily be detected by having the simulation monitor the error associated with the output. The error will indeed be very high whenever the network is presented with new information, that is, questions for which it does not know the answer.

Consider now the case of "can't say" responses. As in the gist condition, the network should produce the answer to the distorted question it is presented with. As the input information is distorted, however, the output will be associated with a higher error than in undistorted cases. This is simply a consequence of the mismatch between the presented and the stored information. Thus, "can't say" responses could also be detected by monitoring the error. But then these responses are indistinguishable from "don't know" responses, because both types entail high error levels. Moreover, if the extent of match is evaluated by the same measure as the quality of retrieval, then we have no way to account for the fact that priming only affects retrieval. Priming can be modeled by giving the network more training on some of the statements in the training set. This would result in strengthened responses to those statements. If the simulation only monitored error, then both ease of retrieval and match would be affected by priming because the training would reduce the discrepancy between the target and actual output.[8] However, this is at odds with our results.

The basic problem is thus that the extent of match and the quality of retrieval are essentially confounded if only the error is monitored. We therefore need two different measures of performance: a measure of the quality of retrieval of the answer and a measure of the quality of the match between the presented information and the stored memory structure. Both measures would be associated with their own criterion. In the gist condition, only retrievability would be monitored, leading to "don't know" or "undistorted" answers according to whether the retrieval criterion has been bypassed or not. In the literal condition, the second measure, the extent of match, would also be monitored, leading to "can't say" answers if the match criterion is bypassed. Otherwise, things would proceed as in

[8] Unless the network is trained with noise in the patterns. In that case, and assuming that the task still consists of reproducing the canonical patterns, additional training will have the effect of improving the network's ability to ignore distorted patterns, thus making it harder to detect the illusion.

the gist condition. The longer reaction times observed in the literal condition would be accounted for by the additional processes involved in monitoring this second measure.

The difficult point in this reasoning is to find two measures of the network's performance that behave in accordance with the condition that strengthening the memory traces only affects retrievability. One possibility would be to monitor the error on different subsets of output units. Given that the network we have been considering is a pattern-completion network, the extent of match could be measured by the error associated with those units that reproduce the question itself, whereas retrieval could be evaluated by the error associated with the units representing the answer. In the first case, we measure how consistent the current input is with its stored representation, whereas in the second case we measure how well the answer to the question is produced by the network. These measures seem reasonable and would effectively allow us to generate all the observed responses of the literal task. However, it is not clear whether they satisfy the crucial condition that only retrieval of the answer should be affected by priming. This is an empirical question that needs further exploration.

C. Implications for Further Research

What are the major implications of this study? First, it seems clear that comprehension in general is expectation driven and highly tolerant of degraded input. As was pointed out earlier, these features do not stem from our willingness to be cooperative, but, rather, reflect central properties of the cognitive system. This view seems intuitive, yet is at odds with research suggesting that people process every word that is read (Just & Carpenter, 1987). Second, comprehension appears to be highly flexible in that the tolerance to degraded input may be controlled strategically. Finally, information retrieval and evaluation of matches seem to be dissociated to some extent. Our attempt at sketching a possible model for the various aspects of the Moses effect suggests that the approach might well shed new light on this phenomenon.

Acknowledgments

The work reported here was sponsored by Grants BNS-003711 and IRI-8719469 from the National Science Foundation, in part by the Office of Naval Research, Contract No. N00014-84-K-0063, Contract Authority Identification Number NR667-529 to the first author. The second author was supported by a fellowship from the Belgian American Educational Foundation. We thank C. Dennler, J. Martin, G. Wells, and C. Wible for help with various aspects of the experiments.

References

Bredart, S., & Modolo, K. (1988). Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica*, **67**, 135–144.

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

Cohen, J. D., Dunbar, K., & McClelland, J. L. (in press). *On the control of automatic processes: A parallel distributed processing model of the Stroop effect. Psychological Review.*

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

Fodor, F. D., & Frazier, L. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, **6**, 291–325.

Frazier, L. (1987). Theories of sentence processing. In J. Garfield (Ed.), *Modularity in knowledge representation and natural language processing*. Cambridge, MA: MIT Press.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **11**, 702–718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory and Cognition*, **15**, 84–93.

Glenberg, A. M., Sarocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, **116**, 119–136.

Graesser, A. C., & Murachver, T. (1985). Symbolic procedures of question answering. In A. C. Graesser & J. B. Black (Eds.), *The psychology of questions*. Hillsdale, NJ: Erlbaum.

Grice, L. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Ed.), *Syntax and semantics: Speech acts* (Vol. 3). New York: Academic Press. Originally published from William James Lectures, Harvard University.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

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

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Newton, MA: Allyn and Bacon.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363–394.

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

MacWhinney, B., Pleh, C., & Bates, E. (1985). The development of sentence interpretation in Hungarian. *Cognitive Psychology*, **17**, 178–209.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

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

Ratcliff, R., & McKoon, G. (1989). Similarity information vs. relational information: Differences in time course of retrieval. *Cognitive Psychology*, **21**, 139–155.

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

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Singer, M. (1984). Toward a model of question answering: Yes–no questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 285–297.

Singer, M. (1985). Mental operations of question answering. In A. C. Graesser & J. B. Black (Eds.), *The psychology of questions*. Hillsdale, NJ: Erlbaum.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, **223**, 30–36.