JOURNAL OF MEMORY AND LANGUAGE 28, 547–563 (1989)

# Assessing the Occurrence of Elaborative Inference with Recognition: Compatibility Checking vs Compound Cue Theory

GAIL MCKOON AND ROGER RATCLIFF

*Northwestern University*

In previous research, we have used delayed item recognition to investigate the information about elaborative inferences that is generated during reading. Recently, Potts, Keenan, and Golding suggested that results from recognition experiments are not determined by information encoded during reading, but rather by information calculated at the time of the recognition test. They proposed that test items that represent potential inferences from studied sentences are compared to the sentences at the time of the recognition test, and performance on recognition is a function of the compatibility of the test items with the sentences in memory. In this article, we present both theoretical and experimental results that argue against this compatibility-checking hypothesis. In Experiment 1, on-line lexical decision is used to demonstrate that relations between sentences and test items that affect on-line performance do not affect performance in delayed recognition, as they should according to the compatibility hypothesis. In Experiment 2, it is shown that subjects' ratings of compatibility do not predict recognition performance. It is concluded on the basis of the experimental results that the compatibility hypothesis does not account for delayed recognition performance. In addition, a review of theories of recognition memory shows that the compatibility hypothesis is not consistent with the temporal dynamics of processes in these theories. © 1989 Academic Press, Inc.

Elaborative inferences involved in reading have been studied extensively in the past few years, with attention given to both methodological and theoretical issues. In our research (McKoon & Ratcliff, 1986, 1988), we have claimed that some elaborative inferences are encoded during reading (perhaps minimally encoded), and we have used data from speeded item recognition experiments to support this claim. However, Potts, Keenan, and Golding (1988) recently argued that the recognition paradigm does not measure inferences encoded during reading, but instead inferences made at the time of the recognition test. In this article, we present experimental results that contradict predictions from their argument and support our view that recognition can be used to examine inferences in the mental representation of a text. But we also show how the relation between recognition tests and encoded meaning is more complex than would be indicated by the simple question, "Is an inference encoded or not?"

To explain the hypothesis proposed by Potts et al. to account for recognition performance, the example in Table 1 will be used. There are two sentences, one *predicting* and one *control*. The predicting sentence allows the inference that the actress died as a result of the fall. This inference is an elaborative inference in the sense that, although it adds information to the sentence, it is not necessary to connect the explicitly stated information of the text into a coherent structure. The control version of the sentence contains all the same words that might be semantically related to *dead* (*fell, 14th story*), but does not lead so directly to any inference about death.

This research was supported by NSF grant 85-16350 to Gail McKoon and NSF grant 85-10361 to Roger Ratcliff. We thank Jan Keenan and George Potts for discussion of the research presented in this article, and Pat Carpenter, Ed Shoben, and several anonymous reviewers for suggestions about earlier versions of the article. Correspondence concerning the article and reprint requests should be addressed to Gail McKoon, Psychology Department, Northwestern University, Evanston, IL, 60208.

TABLE 1
DATA FROM MCKOON AND RATCLIFF (1988b): PERCENT ERRORS ON TARGET WORDS

*Study predicting sentence*
The director and the cameraman were ready to shoot closeups when suddenly the actress fell from the 14th story.

*Test*

| Prime | ready | actress |
|---|---|---|
| Target | dead | dead |
| Errors | 34% | 48% |

*Study control sentence*
Suddenly the director fell upon the cameraman, demanding closeups of the actress on the 14th story.

*Test*

| Prime | ready | actress |
|---|---|---|
| Target | dead | dead |
| Errors | 23% | 21% |

To examine these kinds of inferences, we have used a study–test recognition procedure. Subjects read several unrelated sentences in a study list, and then, in an immediately following test list, pairs of words are presented for test. The first word of the pair is a prime word, and subjects make no response to it. It can be either a neutral word (e.g., *ready*) or a word from a studied sentence. The second word is the target, and subjects are required to decide whether or not it appeared in any of the sentences just studied. Some of the targets that appear in a test list are words from the sentences and so the correct response for them is "yes." Other targets are words totally unrelated to any sentence, and the correct response is "no." The interesting targets are those that express potential inferences. For both the predicting and the control sentences in Table 1, the correct response to *dead* as a target should be "no," because *dead* did not appear in the sentences. But subjects who read the predicting sentence might make more errors on *dead* than subjects who read the control sentence, because of the relation between *dead* and the predicting sentence.

To demonstrate that the increased errors result from information encoded during reading and *not* inference processes that occur when the target *dead* is presented for test, one requirement is to show that the recognition retrieval processes are automatic and do not involve slow, conscious strategies. This can be done by applying the criteria for automatic processes proposed by Posner and Snyder (1975). First, in our experiments, the time available for retrieval is kept too short for strategic processing. The prime-to-target SOA (stimulus onset asynchrony) is 200 ms or less, and subjects are required to respond at an experimenter determined deadline that limits response latencies to 600 ms. Second, the probability that a prime and target are related to each other (via elaborative inference) is kept low, so that there is no reason for subjects to make up an appropriate strategy, even if they had time to execute it. Using these two criteria, we have argued that performance in recognition is not the result of strategic processing (McKoon & Ratcliff, 1986; see also Ratcliff & McKoon, 1981), but that the recognition processing is automatic. In addition, we have argued that automatic recognition processing reflects the interaction of the test item with information encoded during reading.

Potts et al. (1988) accepted that the recognition processes involved in speeded recognition are automatic, but proposed a retrieval process that is not based on encoded inferential information. Their suggestion was designed to explain a specific pattern of data, shown in Table 1 (taken from McKoon and Ratcliff, 1988b; replicated in McKoon & Ratcliff, 1986). In these data, error rates are significantly increased for the target word when the predicting sentence was studied, but only when the target is primed by a word from the studied sentence and not when the prime word is neutral. So with the prime from the sentence, there are more errors in the predicting condition than the control condition (48% versus 21%), but with the neutral prime, the difference is much smaller (34% versus 23%). We interpreted this pattern as showing a minimal or partial inference; the predicted word *dead* is not encoded to a high enough degree to make contact with the mental representation of the sentence on its own, but it is encoded to a sufficient extent that when it is combined with the prime *actress*, it matches the encoded mental representation enough to give errors.

Potts et al.'s interpretation of the data in Table 1 is different. They suggest that no aspect of the inference is encoded at all; instead, the compatibility between the target word and the presented text is computed at the time of the recognition test. The more compatible are the target and text, the greater the probability of an error on the recognition decision. The fact that the difference between the predicted and control conditions is larger with the prime from the text than the neutral prime (Table 1) is due to the text prime "helping insure that the subject performed the compatibility check on the correct sentence" (Potts et al., 1988, p. 402). Thus, in Potts et al.'s interpretation of the data in Table 1, increased errors result from a compatibility check performed at the time of the recognition test, and not from information encoded at the time the predicting sentence was read.

The compatibility checking idea can account for the data shown in Table 1, but it has difficulties with the data shown in Table 2 (from McKoon & Ratcliff, 1988d). The procedure used to obtain these data was speeded item recognition, as described above and the same procedure that was used for the data of Table 1. Again, there were two kinds of sentences, predicting and control. The predicting sentences were written to point to one, highly typical, exemplar of a contextually defined category; for example, the most likely exemplar for the category *animals milked on farms* is *cow*. They were also written to have a secondary exemplar (*goat* in the example). As shown in Table 2, when the two exemplars are used as target words, they give very different results. For the target word *cow*, there are significantly more errors for the predicting than the control sentences, both with the prime from the sentence and with the neutral prime. For the target word *goat*, there are not significantly more errors for the predicting than the control sentences, in either priming condition. We have interpreted this difference in the pattern of results as reflecting a relatively high degree of encoding for information about the typical exemplar in contrast with minimal (or nonexistent) encoding of information for the secondary exemplar (McKoon & Ratcliff, 1988d). The high degree of encoding for the typical exemplar is assumed to be the result of the availability during reading of long-term memory associations that relate the meaning of the predicting sentence to the exemplar (see McKoon & Ratcliff, 1988b, 1988c, and 1988d for further discussion). These findings were replicated conceptually by experiments that show inferences about predictable events are encoded to a higher degree when the sentences describ-

TABLE 2
DATA FROM MCKOON AND RATCLIFF (1988d): PERCENT ERRORS ON TARGET WORDS

*Study predicting sentence*
The old man loved his granddaughter and she liked to help him with his animals; she volunteered to do the milking whenever she visited the farm.

*Test*

| Prime | ready | granddaughter |
|---|---|---|
| Target | cow | cow |
| Errors | 66% | 63% |
| Prime | ready | granddaughter |
| Target | goat | goat |
| Errors | 41% | 38% |

*Study control sentence*
The old man loved his granddaughter and she liked to help him with his animals when she visited the farm; she also liked the milk and cookies her grandmother provided.

*Test*

| Prime | ready | granddaughter |
|---|---|---|
| Target | cow | cow |
| Errors | 40% | 51% |
| Prime | ready | granddaughter |
| Target | goat | goat |
| Errors | 41% | 36% |

ing the events use words semantically associated to the events (e.g., using the words *water, colder, winter,* and snow to predict *freezing,* McKoon & Ratcliff, 1988b).

Both the data for the most typical (primary) exemplar and the data for the secondary exemplar present problems for the compatibility checking hypothesis. For the primary exemplar, compatibility checking cannot explain the data because the error rate is high with the neutral prime even though there is no cue to ensure compatibility checking on the appropriate sentence in memory. And for the secondary exemplar, compatibility cannot be operating because there is no increase in error rates for the predicting sentence relative to the control sentence. However, it may be that the compatibility checking hypothesis does not apply for these materials. For the primary exemplar, it would not apply if the exemplar is strongly encoded into the mental representation of the predicting sentence. Compatibility checking would only come into operation for targets that were not encoded during reading. And for the secondary exemplar, it could be argued that it is not compatible enough with the predicting sentence to give increased error rates.

To summarize, Potts et al.'s notion of compatibility checking applies to the pattern of data shown in Table 1; increased errors on predictable targets come from a compatibility check between the target and a mental representation of the sentence cued by the prime from the sentence. The compatibility notion would at first seem to have problems with the data in Table 2, but it could be assumed that compatibility checking does not apply to strongly encoded inferences or to target words that are not compatible with their studied sentences.

In this article, the compatibility checking explanation of delayed recognition data is challenged in three ways. First, it is argued theoretically that there is no current model by which compatibility checking might operate and still be consistent with available research on recognition processes. Second, Experiment 1 shows that the secondary exemplars discussed above, which must be assumed not compatible with their predicting sentences by the compatibility checking hypothesis, are in fact compatible by one measure of compatibility, facilitation in online lexical decision. Third, in Experiment 2, it is shown that subjects' ratings of compatibility do not appropriately predict performance on delayed recognition.

## Theoretical Accounts of Compatibility Checking

In this section, current models of the processes involved in recognition are considered as potential bases for implementations of compatibility checking. The critical claim of the compatibility checking notion (for explaining the data of Table 1) is that errors in recognition are increased with a prime from a predicting text because the prime increases the probability that the appropriate text will be available at the time of the recognition test. Because the text is more likely available, it is more likely that compatibility checking of target against text will be carried out, and so it is more likely that the recognition decision will be incorrect. Following this reasoning, each of the models to be discussed is considered as to how it would provide a mechanism by which a prime would make available the appropriate text and how it could allow computation of the compatibility between the target and the text.

As mentioned above, the processing in the delayed recognition task used in these experiments can be labeled automatic, and not strategic. Thus, the relevant models for consideration are those that attempt to explain the automatic process involved in recognition. The models considered here are based either on spreading activation or compound cue explanations of priming effects.

*Spreading activation models.* One spreading activation model directly relevant is ACT* (Anderson, 1983). This model could be applied to the data for the predicting text in Table 1 as follows: First, the prime from the text would activate all the concepts of the text, making all these concepts available as though the text were in working memory (but with less strength due to forgetting or failure to have encoded some information). As a result of activation by the prime, inferences could be computed from the text and, when the target word was presented, it could be matched against these inferences. If the target word matched a constructed inference, then a correct negative decision would be difficult and an increased error rate would result.

The problem with this account is that the processes by which ACT* might compute inferences are too slow to make them available in the time required (i.e., the 200 ms SOA between prime and target plus about 650–700 ms response time). Retrieval in ACT* would take place through interactions between the prime and elements in memory, via a spreading activation process. Then productions using general knowledge would operate on the retrieved representation of the sentence plus the prime to produce inferences, and the target would be matched against these inferences. This sequence of operations could not happen quickly enough within the parameters of the ACT* framework (see, for example, the estimates of production firing times, a few hundred milliseconds, in Pirolli & Anderson, 1985). So, although ACT* at first seems a likely candidate model with which to implement compatibility checking, the actual parameters of processing time in the model rule it out. Therefore, because ACT* has no mechanisms that could generate inferences in the required time at test, ACT* would have to assume that inferences were encoded during reading of the predicting sentences.

Another possible spreading activation account for compatibility checking would make use of mechanisms suggested by Forster (1981) to explain the processes of integrating words into their contexts during reading. According to Forster's model, as each word is read, it must be checked to determine whether it can plausibly fit into the context of its sentence and prior discourse. This process might also be used in recognition. Spreading activation from the prime word would make available its text, but no inferences would be constructed. Instead, when the target word was presented, part of the process of reading it would be to check it against its context and calculate its plausibility. When the context was the predicting text (just activated by spreading activation), plausibility would be high and errors in recognition would result.

The problem with this mechanism for compatibility checking is that current data suggest the processes involved would be too slow to affect recognition in the required time. If Forster's context checking process were the mechansim responsible for the data of Table 1, then it would also have to explain the data from a recent study of the timing of elaborative inferences (McKoon & Ratcliff, 1988b). In this study, predicting and control sentences were presented to subjects one word at a time, and immediately after the final word of a sentence, a target word was presented for an immediate recognition test. The predicting sentences were of the same kind as those in Table 1; each predicted some event expressed by the target word. However, the prediction of the target word was allowed only by the final word of the sentence. For example, one sentence was *The diver prepared to do a double somersault into the pool; he jumped, spun, and hit the cement;* the target test word was *hurt*. Each word of the sentence was presented for only 250 ms, so the time for computing an inference to connect *hurt* to the meaning of the sentence was 250 ms plus whatever part of the response time for the test word could be used for further processing (the average response time was 768 ms). With this small amount of time, there was no significant difference between the predieting and the control conditions either in response time

or error rate. Although the inferences were computed, as shown by errors in delayed recognition, they could not be computed quickly enough to affect responses in the immediate test. Thus, the conclusion is that the calculation of an inference (or a context checking process) that relates *hurt* to the predicting sentence must take longer than the time given in this experiment (McKoon & Ratcliff, 1988b; see also Till, Mross, & Kintsch, 1988). Assuming that these inference processes are similar to those used in compatibility checking in delayed recognition, then they would be too slow. If the inference processes could not work in time to affect responses in the immediate recognition experiment, then they could not work in time to affect delayed recognition via context checking.

There is also other data to suggest that the relation between a predicted test word and a retrieved representation of the predicting sentence could not be calculated in the 600–650 ms of a speeded recognition decision. In a series of articles, it has been shown that early in the time course of retrieval for single proposition sentences, only information about independent items (e.g., single words) is available. Information about the relations between items, information of the kind that would be used in generating inferences, is not available until 600 to 700 ms of processing has elapsed. This pattern of a delay in the availability of relational information has been shown to apply to several kinds of relations, including agent versus object relationships in simple sentences (Ratcliff & McKoon, 1988c), paired associate relations for words (Dosher, 1984; Gronlund & Ratcliff, 1988), and set inclusion relations for categories (Ratcliff & McKoon, 1982). If these kinds of information are not available as early in processing as 650 ms, then it seems reasonable to assume that inferences would not be available that early either. Whether the inferences were generated from the representation of the text activated by the prime word, or the relation between the target word and the representation of the text were calculated when the target was presented, there would not be sufficient time.

Up to this point, two ways of implementing compatibility checking have been considered, both based on spreading activation. In the first (ACT*), the prime from the predicting text was assumed to activate the respresentation of the text in memory, and then an inference about the predicted event was generated from this representation. In the second, spreading activation was also assumed to activate the representation of the text, but the connection between the predicted test word and the text was assumed to be calculated (via context checking) at the time the test word was presented. For both of these ideas, the problem is that the processes that would generate the inference or relate the text and target word would be too slow to affect speeded recognition decisions in the experiments of Tables 1 and 2.

*Global memory models.* An alternative to spreading activation is the resonance notion of some current memory models (Gillund & Shiffrin, 1984; Hintzman, 1986; Murdock, 1982; Ratcliff, 1978, 1988). According to these models, a target word (or cue) is compared against information in memory by a passive, parallel process that yields a measure of the target's overall familiarity or goodness of match. This goodness of match represents the interaction between encoded information in memory and the target, and the value of the match determines the speed and accuracy of the response to the target (Ratcliff, 1978). To account for priming phenomena, these models assume cue-dependent retrieval (Craik & Tulving, 1975; Gillund & Shiffrin, 1984; Tulving, 1974); that is, they assume that the probe to memory is not the target alone but rather the target in its retrieval context. When a target is preceded by a prime, the target and prime are assumed to combine into a compound cue, with the prime weighted less than the target (Ratcliff & McKoon, 1988a, 1988b). The familiarity



(goodness of match) of this compound determines the response to the target. The account of the data of Tables 1 and 2 that would be given by these models is that prime-target compounds are compared against memory, and a high value of goodness of match for an inference target leads to a high error rate. When the value is high for an inference target (e.g., *cow,* Table 2), then there must be some information relevant to that target in the mental representation of the text. Furthermore, because the value of match depends on the compound of prime and target, the value can be different with different primes (e.g., *dead,* Table 1). This account of the data is discussed in detail by Ratcliff and McKoon (1988a), where implementation in a number of models is presented and supporting data is reviewed. In the present context, this compound cue notion, based on cue-dependent retrieval, accounts for data of the kind shown in Tables 1 and 2 by assuming that speeded recognition responses reflect an interaction of the prime-target compound with information encoded into memory during the reading of study texts.

To summarize, the only current models by which inferences can affect speeded recognition decisions place the inferences in the representation of the text that was constructed during reading. ACT* must make this assumption in order to retain time course parameters that allow it to account for other data. The global memory models make this assumption as part of their commitment to passive parallel mechanisms for recognition. Furthermore, data about the time course of retrieval processes rule out the construction of inference relations in the time required for a speeded recognition decision. For all of these reasons, compatibility checking is an unlikely explanation of the data.

The argument that compatibility checking is an unlikely mechanism because it is inconsistent with theories may seem to be a weak argument because any current theory is almost certainly wrong. However, the strength of the argument lies not only in the theories themselves but also in the large range of data they represent. ACT* has been used to account for a variety of data from different domains (Anderson, 1983). The global memory theories (Gillund & Shiffrin, 1984; Hintzman, 1986; Murdock, 1982; Ratcliff, 1988; Ratcliff & McKoon, 1988a) have explained data from recognition, recall, frequency judgments, categorization, and various reaction time paradigms. By keeping our explanation of recognition data broadly consistent with the theories, the explanation is also automatically consistent with all of this data. The compatibility checking hypothesis, on the other hand, has not been shown to be consistent with any data other than that for which it was originally formulated.

If the compatibility checking hypothesis is wrong, as the theoretical arguments suggest, then it should also be possible to argue against it empirically. We do this in the next sections of this article.

## Experiment 1

In order to give a compatibility checking account of the data in Table 2, it must be assumed that the secondary exemplars (e.g., *goat*) are not compatible enough with their predicting texts to affect a recognition decision. The compatibility checking account would be called into question if it could be shown that the secondary exemplars were, in fact, compatible with their predicting sentences. There would be several possible ways in which compatibility might be demonstrated. For example, in Experiment 2, subjects' ratings were collected. However, in Experiment 1, we wanted a task that would tap processes more similar to those that might be proposed to occur in compatibility checking. The task we chose was on-line lexical decision.

In Experiment 1, subjects read sentences like those in Table 2, and each sentence was followed by a string of letters presented for lexical decision. The sentences of interest

were presented in either their predicting or control versions, and they were followed by a target word that expressed one of the possible inferences, either the primary exemplar (*cow*) or the secondary exemplar (*goat*). Other, filler, sentences were used to precede targets that were nonwords.

On-line lexical decisions are often claimed (cf. Forster, 1981; Potts et al., 1988) to be based (at least in part) on compatibility or context checking from the target word back to the text. If such checking occurred in Experiment 1, and if the secondary exemplars were compatible with their predicting sentences, then facilitation ought to be observed in the lexical decisions on the secondary exemplars.

Of course, facilitation in on-line lexical decision is not necessarily due to compatibility checking; it may also be due to priming of lexical access. However, whatever the process responsible for the facilitation, the force of the argument remains the same. According to Pott's et al.'s reasoning, the same processes that determine the relation between the target and the sentence in the on-line task should also operate in the delayed recognition task when the target in delayed recognition is preceded by a prime from its text. This is because, by their argument, the prime serves to make the text available just prior to presentation of the target.

In sum, the prediction from the compatibility hypothesis is that the relation between a secondary exemplar and its predicting text should have the same effect for on-line lexical decision as for delayed recognition. If the compatibility between the exemplar and the text is so low that it does not affect delayed recognition (as is the case in Table 2), then it should also not affect on-line lexical decision.

### *Method*

*Materials.* The set of 32 sentences was the same as the set used in McKoon and Ratcliff (1988d). Each sentence had a predicting and a control version and each sentence mentioned some category (e.g., *animals* in the sentences in Table 2). The predicting version of the sentence pointed to particular members of the category and was consistent with more than one category member. The target words were one category member that was highly related to the predicting sentence (*cow* in Table 2) and one that was less highly related (*goat* in Table 2).

There were also 32 filler paragraphs of the same general style, and a list of pronounceable nonwords to be used for negative test items.

*Subjects.* Undergraduates participated in the experiment for credit in an introductory psychology course. Two groups of subjects were tested; one group (18 subjects) was tested with the highly related, primary, target words for all experimental materials, and one group (24 subjects) was tested with the less related, secondary, targets for all experimental materials.

*Procedure.* All stimuli were presented on a CRT screen, and responses were collected on the CRT's keyboard. The CRT was driven by a real-time computer system that controlled presentation times and measured response times.

For practice, subjects were given a list of 25 lexical decision test items without preceding sentences. Each item was preceded by a warning signal of three asterisks presented for 250 ms, then the item was displayed until a response key was pressed, and then, after a 1 s pause, the next item's warning signal was presented. Subjects were instructed to respond quickly and accurately, using their right index finger on the keyboard's "?/" key for a positive (*word*) response and their left index finger on the "Z" key for a negative (*nonword*) response.

After the lexical decision practice, subjects were given 10 practice sentences, and then the 32 experimental and 32 filler sentences in the experiment proper. Presentation of each sentence began with a message displayed on the CRT instructing the subject to press the space bar on the keyboard; then after a 1 s pause, the sentence was displayed. The sentences were displayed left-justified on the CRT screen and were all two lines long. A sentence remained on the CRT screen until the subject pressed the space bar again. Then, a row of three asterisks was presented, left-justified on the next line of the CRT. The asterisks were displayed for 250 ms, and then the lexical decision test item was presented immediately under the asterisks. The test item remained on the screen until a response key was pressed; then the screen was cleared, and, after a 1 s pause, the message to press the space bar to initiate the next sentence was presented. Subjects were instructed to read the sentences carefully for comprehension, and to respond quickly and accurately on the lexical decision test items, using the same response keys as in the practice.



To encourage reading for comprehension, subjects were given written questions requiring short answers. A list of five questions was given after the 10 practice sentences, and a list of eight questions was given at the end of the experiment. Performance on these questions was not scored.

*Design.* The experimental sentences were presented in their predicting on their control versions. The targets were always the same for a group of subjects; for one group, always the primary target, and for the other group, always the secondary target. The predicting versus control variable was combined in a Latin square design with two sets of sentences (16 per set) and two sets of subjects (9 in each set in the primary target group, and 12 in each set in the secondary target group). Order of presentation of sentences was random, different randomization for each two subjects.

### *Results*

The most interesting question is whether the predicting sentences affected performance on lexical decisions for the secondary target words. The answer is that they did, with the effect apparent in both error rates and response times, as shown in Table 3. The difference in response times was significant with subjects as the random variable, $F(1,23) = 4.68$, $p < 0.05$, and marginally significant with target words as the random variable, $F(1,31) = 3.84$, $p < 0.10$. The difference in error rates was significant in both analyses, $F(1,23) = 23.3$ and $F(1,31) = 5.7$, $p$'s $< 0.05$. The standard error for response times was 12.6 ms. The average response time for nonwords was 775 ms, with 6% errors.

Not surprisingly, for the primary targets (the second group of subjects), lexical decisions following predicting sentences were faster than lexical decisions following control sentences. This difference was significant with subjects as the random variable, $F(1,17) = 21.3$, and with items as the random variable, $F(1,31) = 17.2$. The standard error for the response time means was 13 ms. Response times for nonwords averaged 768 ms, with 4% errors.

For both groups of subjects, reading times showed a tendency to be faster for the predicting sentences than for the control sentences. However, this tendency did not lead to uniformly significant differences, so it is not possible to draw strong conclusions and the results are simply summarized here. For the group of subjects who had the secondary targets, predicting sentences were read faster than control sentences, 7731 ms versus 8062 ms, but this difference was not significant with sentences as the random variable, $F(1,31) = 1.1$, and was only marginally significant with subjects as the random variable, $F(1,23) = 3.48$, $p <$

TABLE 3
DATA FROM EXPERIMENT 1: RESPONSE TIMES (IN MS) AND PERCENT ERRORS

| | Primary exemplars | | Secondary exemplars | |
|---|---|---|---|---|
| Sentence | RT | % errors | RT | % errors |
| Predicting | 551 | 2 | 661 | 3 |
| Control | 637 | 2 | 690 | 10 |

0.10. Reading times for the filler sentences averaged 6523 ms.

Reading times for the subjects with the primary targets were faster for the predicting sentences than the control sentences, 7061 ms versus 7488 ms. This difference was not significant with sentences as the random variable, $F(1,31) = 2.45$, but was significant with subjects as the random variable, $F(1,17) = 5.7$, $p < 0.05$. Reading times for the filler sentences averaged 5761 ms.

*Discussion*

The results of Experiment 1 are not consistent with the compatibility hypothesis. If, in on-line lexical decision, the relation between the secondary target and the predicting text affects responses, then this relation should also affect responses in delayed recognition. A prime from the predicting text should make the text available, and then the relation between the predicted target and the text should affect processing similarly to the way it affects processing when the text is immediately available in the on-line task. But this is not what the data show; Experiment 1 shows that the relation affects on-line lexical decisions, whereas Table 2 shows that there is no significant effect in delayed recognition.

There is one potential problem that could be raised with the interpretation of the results of Experiment 1 and Table 2, and that is that the power of the two tasks, lexical decision and recognition, may be different. However, this is not the case. A comparison of the power of the two experimental procedures shows that the power in recognition is at least as great as the power in lexical decision. In recognition, for the primary exemplars, the average error rate for the predicting condition was 8.0 standard errors greater than the average rate for the control condition. This was larger than the difference for response times in lexical decision, where in the predicting condition average latency was 6.8 standard errors faster than in the control condition. Thus, in terms of standard errors, recognition appears to be somewhat more powerful than lexical decisions, not less powerful. So the lexical decision difference between the predicting and the control conditions for the secondary exemplars, a difference of 2.5 standard errors in latencies (and a larger difference in error rates), should be observable in the recognition data. The fact that it is not indicates that whatever the relation between the secondary target and its predicting sentence that affects on-line lexical decision, this relation does not affect delayed recognition. Thus, the results of Experiment 1 are not consistent with the compatibility hypothesis.

In contrast, the compound cue notion (Ratcliff & McKoon, 1988a) can account for both the results shown in Table 2 from delayed recognition and the results of Experiment 1. In Table 2, the error rate for the primary exemplar is high after the predicting text because information about the primary exemplar was encoded during reading of the text. The error rate for the secondary exemplar is not significantly higher after the predicting than the control text because information relevant to it was not encoded during reading. These data were generated in a situation where the compound cue (made up of the information in short-term memory) is the target and its prime (and perhaps some small contribution from the previous test item). In Experiment 1, the situation is different. When a target is presented immediately after a sentence, the contents of short-term (or working) memory are the words in the sentence plus the propositions (meanings) formed out of those words. These kinds of information combine with the target to form the compound that determines the familiarity of the target (see Ratcliff & McKoon, 1988a for further discussion). Both the primary and the secondary targets combine with the predicting sentence to form familiar compounds (both *goats* and *cows* are animals that can be milked on farms), and so responses to both are facilitated. The critical point is that the secondary target can form a familiar compound when it is actually presented in conjunction with the predicting text, even though in delayed recognition, it does not match the information encoded from the text alone. Thus the compound cue theory can account for both the on-line data and the delayed recognition data.



## Experiment 2

Experiment 2 was designed to test the compatibility hypothesis in a different way from Experiment 1, by determining whether subjects' compatibility ratings would predict performance in delayed recognition. Two sets of materials were used in the experiment. The first was the set shown by example in Table 1 (and used by McKoon & Ratcliff, 1988b), in which each predicting sentence predicts some event represented by the test word. These will be referred to as the *predicting event* materials. The second set is shown by example in Table 2 and was used in Experiment 1 and by McKoon and Ratcliff (1988d). In this set, each predicting sentence was associated with two possible category exemplars, the primary exemplar and the secondary exemplar. These materials will be referred to as the *category exemplar* materials. In the experiment, the sentences from both sets were presented to subjects along with a target word, and the subjects were asked to rate how compatible each target was with its sentence. The sentences were presented either in their predicting or their control versions. The target word for the predicting event sentences was the predicted word, and the target for the category exemplar sentences was either the primary or the secondary exemplar. The compatibility ratings collected in this experiment were then used to test whether they could predict the probabilities of recognition errors for the data shown in Tables 1 and 2 from McKoon and Ratcliff (1988b, 1988d).

According to the compatibility hypothesis, compatibility ratings should predict performance in delayed recognition, as illustrated in Fig. 1. The predictions can be understood in the following way: First, suppose that a prime from the text ensures that inference target words (such as *dead* for the text in Table 1) are checked against the appropriate sentence in memory. Then the probability of an error should increase with compatibility.

Second, suppose that target words are never checked against the appropriate sentence in memory. Then compatibility will have no effect on the probability of a recognition error, and the function relating compatibility and probability of an error will be flat. This would happen for target words without a prime from the appropriate text that were not checked against that text. With two studied texts, the unprimed targets would be checked against the wrong text half of the time, on average, and so for these targets, the function would be flat. For the half (on average) of the unprimed targets that were checked against the appropriate text, the function would be the same as for the primed targets. Therefore, combining the two halves, the slope of the function for unprimed targets should be half the slope of the function for primed targets, as shown in Fig. 1. Both functions start at the same point, a baseline error rate. This is because if there is zero compatibility, the probability of an error should be at some baseline that represents the inherent inaccuracy of the experimental conditions.

According to the compatibility checking hypothesis, errors in recognition for inference target words result from a compatibil-

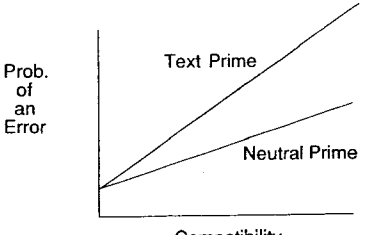


FIG. 1. Predictions of the effect of compatibility on the probability of an error in delayed recognition.

ity checking process that occurs when the target is presented, and so the interaction in Fig. 1 must be predicted. In contrast, the theories that explain recognition errors by assuming that information about the target word is part of the encoding of the studied text (compound cue theory, or versions of ACT*) do not have any obvious way to make predictions about the relation between compatibility ratings and error probabilities. Before the theories could do this, compatibility ratings and recognition responses would have to be linked theoretically. In principle, a direct theoretical link is not possible because compatibility ratings must reflect processing that occurs when the text and the target word are presented at the same time, for as long as the subject wants to consider them, whereas for recognition responses, the target must be matched quickly against a text that was presented by itself without the target. Indirect links between the processes of combining presented pieces of information (the text and the target) into a compound to be matched against memory and the processes of generating inferred information (from the text alone) have not been explicitly considered in any current theory. Thus, empirical functions relating compatibility ratings and probability of recognition error are a test of the compatibility checking view but not of other theories.

### *Method*

*Materials.* There were two sets of materials, the category exemplar materials used in Experiment 1 and by McKoon and Ratcliff (1988d, and the predicting event materials used by McKoon and Ratcliff (1988b). For each of the 32 category exemplar items, there was a predicting sentence and a control sentence, and a primary target and a secondary target. For each of the 28 predicting event items, there was also a predicting and a control sentence, and a target that expressed the event predicted by the predicting sentence.

*Subjects and design.* For the category exemplar materials, there were four conditions, defined by crossing the sentence (predicting or control) with the target word (primary or secondary). These were combined in a Latin square design with four groups of subjects (eight per group) and four groups of materials. For the predicting event materials, there were two conditions; each sentence was presented in the predicting or the control version. These were also combined in a Latin square design with subjects and materials. The 32 subjects participated in the experiment for extra credit in an introductory psychology course.

*Procedure.* Each subject was given the test items in a booklet, with the 32 category exemplar items and the 28 predicting event items each listed in the experimental condition appropriate for the subject. The items were presented in random order, and subjects were allowed unlimited time to complete the booklet.

For each item in the booklet, the sentence was given, followed on the line below by the target word and the digits one through seven. Subjects were asked to rate the compatibility of the target with the sentence, with seven the most compatible.

### *Results*

For the category exemplar materials, the mean compatibility ratings were as follows: for the predicting sentences, compatibility of the primary targets averaged 6.47 and compatibility of the secondary targets averaged 4.38. The corresponding numbers for the control sentences were 2.91 and 2.47. Thus, as expected, the primary targets were indeed rated by the subjects as more related, or more compatible, than the secondary targets. For the predicting event materials, the average rating of compatibility of the target against the predicting sentence was 6.18 and against the control sentence, 2.70. By analyses of variance (one for the category exemplar materials and one for the predicting event materials), all main effects and interactions for these ratings (with both subjects and items as random variables) had $F$ values greater than 20.



The mean ratings for the primary targets from the category exemplar materials (6.47) and the predicting event targets (6.18) suggest that the compatibility checking hypothesis cannot account for the probabilities of recognition errors. The pattern of recognition data for the primary targets (Table 2) and the predicting event targets (Table 1) is much more different than would be expected from the small difference in compatibility ratings. (In fact, the distributions of compatibility ratings for the two sets of materials overlap almost exactly, except that the category exemplar materials have a higher concentration of scores in the higher compatibility range.)

The compatibility checking prediction shown in Fig. 1 was also tested. According to the hypothesis, a prime from the text makes available the text as a whole and then compatibility of the target word can be judged against the text. This hypothesis was proposed to account for performance with the predicting texts, where the error rate increased with a prime from the text relative to the neutral prime. For example, with the predicting event materials, the difference between the predicting and control conditions was increased by a prime from the text from 11% to 27% (see Table 1; data from McKoon & Ratcliff, 1988b). Compatibility differences among items should be able to account for at least a large proportion of this 27% versus 11% difference in recognition performance. In other words, as shown in Fig. 1, the effect of compatibility on recognition with the prime from the text should be much larger than the effect of compatibility with the neutral prime (large enough to account for the ratio 27/11 in recognition performance).

To test this prediction, regression analyses were performed on the compatibility ratings collected in this experiment versus the proportions of errors from the previous experiments (for individual items). The regression analyses of interest for the predicting event materials (Experiment 1; McKoon & Ratcliff, 1988b) are those for the predicting text, with the prime from the text and with the neutral prime. In neither of these conditions was the correlation between compatibility and proportion of errors significant. For the condition with the prime from the text, $R^2 = 0.102$, $F(1,26) = 2.96$, and the slope of proportion correction regressed against compatibility was $0.186 \pm 0.108$. For the neutral priming condition, $R^2 = 0.067$, $F(1,26) = 1.85$, and the slope was $0.124 \pm 0.091$. In neither case was proportion correct significantly predicted by compatibility and the effects are not significantly different from each other. These findings are clearly counter to the compatibility hypothesis.

Regression analyses were also performed on the data from the category exemplar materials (Experiment 1; McKoon & Ratcliff, 1988d). When the analyses were done separately for the primary and secondary targets, the results were about the same as for the predicting event materials, with no significant differences between the condition with the prime from the text and the neutral condition. To give the analyses more power with a broader range of compatibility ratings, the primary and secondary targets were combined. For the prime from the text, $R^2 = 0.29$, $F(1,62) = 25.9$, and the slope was $0.110 \pm 0.022$. For the neutral prime, $R^2 = 0.19$, $F(1,62) = 14.5$, and the slope was $0.096 \pm 0.025$. With this analysis, compatibility does have a significant effect, as would be expected, but the effect is not significantly different for the condition with the prime from the text than the neutral priming conditions; the slopes are within one standard deviation of each other. Again, the results are counter to the compatibility hypothesis; they do not show the interaction in Fig. 1.

Overall, the probabilities of recognition errors are not highly correlated with compatibility ratings. The question then arises as to what empirically available measure

might give a higher correlation. One possibility is a "predictability" measure, asking subjects to rate how "predictable" a target word was in the context of a preceding sentence. However, this measure does not adequately predict recognition errors (McKoon & Ratcliff, 1988b). Another possibility is to present the sentence alone, without the target word, and ask subjects to write "what will happen next." This measure has the advantage that it matches the study conditions for recognition in that the sentence is presented alone. However, this measure also fails to predict recognition errors (McKoon & Ratcliff, 1988b).

The reason that these two measures fail to predict recognition errors may be that they do not tap the meanings of sentences that are encoded during reading, the meanings that would be encoded when the sentences are presented for study in the recognition experiments. The words of a sentence combine and interact to focus on specific aspects of meaning (McKoon & Ratcliff, 1988a). A sentence with several words that are semantically associated to a target may be encoded with one aspect of its meaning focused on that target. For example, a sentence that predicts the target word *sew* and has in it the words *seamstress, thread,* and *needle* appears to have information about *sew* encoded as part of its meaning (McKoon & Ratcliff, 1988b). A sentence that contextually defines a particular category (*animals that are milked on farms* may have as part of its meaning a focus on a specific exemplar of that category (*cow*) (McKoon & Ratcliff, 1988d). How to measure these "focused" aspects of meaning with a task other than recognition is a task yet to be undertaken.

## Discussion

The purpose of this article is to challenge an hypothesis put forward by Potts et al. (1988) about the mechanism responsible for performance in recognition tasks investigating elaborative inferences. They argued that a prime from a studied sentence functions as a cue to activate the sentence, and then the compatibility of a target word is judged with respect to the activated sentence. If a target word representing an inference is compatible, then a correct negative decision on the target will be difficult, and errors will result. By this mechanism, errors in recognition do not show information encoded at the time of reading the sentence, but rather the results of compatibility checking performed at the time of the recognition test.

This hypothesis is challenged in this article in three ways. First, it is argued that there is no current theory of recognition performance that can provide both a mechanism for compatibility checking and still account for the wealth of data about retrieval processes in memory. While these theories are certainly far from complete, they share a common core of assumptions by which they capture a large range of data; it is this communality that we contrast to compatibility checking. Second, in Experiment 1, on-line lexical decisions for secondary category exemplar targets were faster and more accurate following predicting than control sentences. This result predicts, by Potts et al's proposal, that decisions in delayed recognition should also be affected by the predicting versus control sentence variable. A prime from the sentence should activate the sentence, and then the relation between the sentence and the target should affect recognition responses, by the same mechansim that affects lexical decision in the on-line task. In fact, however, the relation between the sentence and the target did not affect recognition responses for the secondary targets. Third, according to the compatibility checking hypothesis, compatibility ratings should predict probabilities of recognition errors. But targets with about the same compatibility ratings showed quite different patterns of recognition errors. In addition, the hypothesis predicts a larger correlation between compatibility ratings and recognition errors with a prime from the text than with a neutral prime, but the data showed no differences among the priming conditions. We conclude, therefore, that performance in speeded item recognition does not depend on a compatibility check between target words and texts. Instead, we follow current memory theories in assuming that recognition decisions depend on a passive, parallel matching process that compares a test item to information in memory. By these theories, recognition responses are based on an interaction between the test item, in its retrieval context, and encoded information. Thus, speeded item recognition is a valid method of investigating the products of the elaborative inference processing that occurs during reading.



Despite our disagreement with Potts et al. (1988) about the use of recognition, we do think they are right about one important point. They say that, with a recognition task, "in order to decide whether a word was presented in a text or not, it is necessary to examine the memory representation of the text." We think this is exactly as it should be. A task that does not force examination of the representation of the text may not show what information is in that representation. For example, Potts et al. propose to investigate inference processing by measuring naming latencies on inference test words (e.g., measuring naming latency for *dead* following predicting or control sentences as in Table 1). They assume, implicitly, that if a concept is inferred then there will be activation in the lexicon of information that is used for naming that concept. However, according to several models (Cottrell & Small, 1983; Fodor, 1983; Norris, 1986; Seidenberg, Dunbar, & Cohen, 1988; see also McKoon, 1988), information at the textual level should not influence information at the lexical level. Potts et al. argue against these models with data that show one circumstance in which textual information does appear to affect naming latency. However, in this instance, the materials may have specified the lexical features of the test words so exactly that the concepts might have been primed for naming through lexical information and not textual information, and so the result would be consistent with models that separate textual from lexical information (Seidenberg et al., 1988). Thus, there is no reason to think that naming latencies can measure inference processes, unless the processes point very specifically to a particular lexical item.

Potts et al. point out that "attempts to determine whether inferences are drawn while reading have been plagued by the inability to unambiguously determine whether the results reflect the operation of processes that occurred while reading or processes that occurred at the time of test." In our view, one reason for the plague is that the question is wrong. In recognition, automatic retrieval processes that are based on information in memory *cannot* be separated from encoding processes, because retrieval is cue-dependent (Tulving, 1974; see also McKoon & Ratcliff, 1988a; Ratcliff & McKoon, 1988a; 1988b). This point is demonstrated by the results in Tables 1 and 2. The question of whether *dead* is encoded as part of the mental representation of its predicting sentence would have one answer ("yes") for retrieval of *dead* with the cue *actress,* and a different answer ("no") for retrieval of *dead* with the neutral prime. In comparison, *cow* acts as though it is part of the mental representation of its predicting sentence under two retrieval conditions (a cue from the predicting sentence and the neutral cue). Perhaps if *cow* appeared to be part of the mental representation under a sufficiently large number of retrieval conditions, we would want to conclude that it was actually explicitly encoded. But we can at least conclude on the basis of results currently available that it is encoded to a higher degree than *dead,* because it is retrieved under a wider variety of retrieval conditions.

In conclusion, results from the speeded recognition task that we have used to examine elaborative inferences are not due to a compatibility checking process that depends only on information generated at the time of the recognition test. Contrary to the suggestion of Potts et al. (1988), the results do allow conclusions about the relation between inferences and the encoded mental representation of a text. However, the results cannot answer the question of whether or not an inference is explicitly encoded, but only indicate under what retrieval conditions evidence for an inference appears. We think that one promising direction for progress in understanding the inferences made during reading lies in research designed to compare different kinds of inferences across a variety of retrieval contexts.

## References


ANDERSON, J. R. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

COTTRELL, G. W., & SMALL, S. (1983). A connectionist scheme for modelling word sense disambiguation. *Cognition and Brain Theory,* 6, 89–120.

DOSHER, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed–accuracy study. *Cognitive Psychology,* 16, 519–555.

FODOR, J. A. (1983). *The modularity of mind.* Boston: MIT Press.

FORSTER, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology,* 33, 465–495.

GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review,* 19, 1–65.

GRONLUND, S. D., & RATCLIFF, R. (In press). The time-course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

HINTZMAN, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review,* 93, 411–428.

MCKOON, G. (1988). Word identification and elaborative inference: The relation between theory and empirical measurement, submitted.

MCKOON, G., & RATCLIFF, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 12, 82–91.

MCKOON, G. & RATCLIFF, R. (1988a). Contextually relevant aspects of meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14, 331–343.

MCKOON, G. & RATCLIFF, R. (1988b). Semantic association and elaborative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 15, 326–338.

McKoon, G. & RATCLIFF, R. (1988c). Dimension of Inference. *The Psychology of Learning and Motivation,* in press.

MCKOON, G. & RATCLIFF, R. (1988d). Inferences about Contextually-Defined Categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* in press.

MURDOCK, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review,* 89, 609–626.

NORRIS, D. (1986). Word recognition: Context effects without priming. *Cognition,* 22, 93–136.

PIROLLI, P. L., & ANDERSON, J. R. (1985). The role of practice in fact retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 11, 136–153.

POSNER, M. I. & SYNDER, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium.* Hillsdale, NJ; Erlbaum.

POTTS, G. R., KEENAN, J. M., & GOLDING, J. M. (1988). Assessing the occurrence of elaborative inference: Lexical decision versus naming. *Journal of Memory and Language,* 27, 399–413.

RATCLIFF, R. (1978). A theory of memory retrieval. *Psychological Review,* 85, 59–108.

RATCLIFF, R. (1988). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review,* 95, 238–255.

RATCLIFF, R., & MCKOON, G. (1981). Automatic and strategic components of priming in recognition. *Journal of Verbal Learning and Verbal Behavior,* 20, 204–215.

RATCLIFF, R., & MCKOON, G. (1982). Speed and accuracy in the processing of false statements about semantic information. *Journal of Experimental Psychology: Human Learning and, Memory,* 8, 16–36.

RATCLIFF, R. & MCKOON, G. (1988a). A retrieval theory of priming in memory. *Psychological Review,* 95, 385–408.

RATCLIFF, R. & MCKOON, G. (1988b). Memory models, text processing, and cue-dependent retrieval. In F. I. M. Craik & H. L. Roediger (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving.* Hillsdale, NJ: Erlbaum.

RATCLIFF, R., & MCKOON, G (1988c). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology,* in press.

SEIDENBERG, M. S., DUNBAR, K., & COHEN, J. (1988). *Lexical semantics and language comprehension.* Paper presented at the Conference on Comprehension Processes in Reading, Wassenaar, The Netherlands.

TILL, R. E., MROSS, E. F., & KINTSCH, W. (1988). Time course of priming for associate and inference words in a discourse context. *Memory and Cognition,* 16, 283–298.

TULVING, E. (1974). Cue-dependent forgetting. *American Scientist,* 62, 74–82.


(Received October 18, 1988)
(Revision received December 2, 1988)