JOURNAL OF MEMORY AND LANGUAGE 31, 591–614 (1992)

Trace Synthesis in Cued Recall

LEIGH ERIK NYSTROM AND JAMES L. MCCLELLAND

Carnegie Mellon University

Several memory models propose that recall may combine traces of different memories. Such models predict blend errors during cued recall. To examine memory blending during recall, four experiments were performed. In each experiment, subjects rated the plausibility of several sentences, many of which shared words with one other sentence. Later, they were asked to recall words from a single sentence to complete partial-sentence cues. When the cue matched two study sentences, subjects made blend errors, recalling one word from each study sentence more frequently than in a control condition. Blend errors were relatively infrequent, however, occurring on about 5% of opportunities. A good account of the results was provided by a stochastic interactive activation model that causes blend errors by synthesizing traces during retrieval.

Many current models of memory account not only for recall and recognition performance, but also for the prototyping and generalization found in concept formation experiments (Knapp & Anderson, 1984; McClelland, 1981; McClelland & Rumelhart, 1985). As a class, such models might be called trace synthesis models, because in all of them the representation retrieved at recall is some form of synthesis of multiple memory traces. Various models differ in their choice of storage representations; some store each trace separately (McClelland, 1981; Hintzman, 1986) while others use superimposed or holographic storage involving matrices (Hinton & Anderson, 1981; McClelland & Rumelhart, 1985; Humphreys, Bain, & Pike, 1989) or vector convolutions (Metcalfe Eich, 1982; Murdock, 1982). Regardless of the form of representation however, retrieval processes in all of these models can lead to prototyping because they involve an activation of more than one trace to produce a set of properties that might not correspond to a single trace as originally stored.

Consequently, it would seem that any of these models would also predict abundant errors during normal recall—*blend* errors, mixing properties from more than one trace. For models which store each trace separately, this synthesis of multiple traces could occur during retrieval, while for models which superimpose traces in memory, the synthesis could occur during storage. It appears that the synthesis-at-retrieval and the synthesis-in-storage classes of models make similar predictions in most circumstances (McClelland & Rumelhart, 1985) and this paper is not intended to distinguish between them.

Not all models of memory retrieval assume that recall involves a synthesis of multiple memory traces, however. Some models view the recall process as a search through memory to find just one single

Preparation of this article was supported in part by a National Science Foundation Graduate Fellowship to Leigh Nystrom, by NIMH Career Development Award MH00385 to James L. McClelland, and by National Science Foundation Grant BNS 88-12048. We thank the members of the first author's research committee, John R. Anderson and Lynne Reder, for their general and specific comments on much of the work presented here. A conversation with Doug Hintzman led to the fourth experiment. We also gratefully acknowledge the PDP research group at CMU for helpful discussions about the model, as well as Michael McCloskey and an anonymous reviewer for their comments on an earlier version of this paper. Correspondence concerning this article and reprint requests should be addressed to Leigh Nystrom, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.

0749-596X/92 $5.00

trace whose properties satisfactorily match a cue. Traces might be viewed as files, with the task of recall as the selection of a file from which to read all the desired information. Most memory models before the 1980s had this character. It would be possible in such a "single-trace" model to synthesize traces during encoding; one could imagine that as someone encoded a new trace, they might have brought older traces to mind and formed a new composite trace relating them all. Note that this limited form of trace synthesis is available in synthesis-at-retrieval or synthesis-in-storage models as well and would be better characterized as an elaboration than as a blend. Because single-trace models do not synthesize multiple traces during either storage or retrieval, they can easily avoid blend errors—though, on the other hand, they do not automatically yield the prototyping or generalization benefits found in the other models.

Shiffrin's SAM model of associative memory (Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984) provides one example of a single-trace model. In the SAM model, cued recall proceeds as a probabilistic selection of a single trace from the collection of all traces. The probability of selecting a given trace is essentially a function of a baseline strength of the trace as well as the degree of match between the cue's properties and the trace's properties. Once selected, all properties of the trace are accessible, and no other traces will be accessed. Another prominent single-trace model is Anderson's ACT theory (1976; 1983), which represents episodically or grammatically related information as being subsumed under a single "trace" node; the goal of recall is the selection of one of the possible trace nodes, as a function of the activation of the nodes subsumed under it. As with Shiffrin's model, once a single trace is selected, all of its constituent information (the subsumed nodes) can be retrieved without interference from other traces. Neither model predicts blend errors.

We can see then that single-trace and trace synthesis models make contrasting predictions about the existence of blend errors during recall. Single-trace models predict that traces should not be blended, because they retrieve each trace as an intact whole. By contrast, synthesis-at-retrieval and synthesis-in-storage models seem to predict an abundance of blend errors, because they fail to explicitly keep each trace intact during the recall process. The intent of this paper, therefore, is not to separate synthesis-at-retrieval from synthesis-in-storage models. For now, we will set synthesis-in-storage models aside, and return to them in the General Discussion. Rather, the purpose of this paper is to examine the phenomenon of blend errors during recall in order to assess the validity of predictions from the synthesis-at-retrieval account; to explore whether a coherent fit to a body of data can be obtained from a model of that type; and to examine whether and to what extent the data are consistent with the other types of models.

The particular synthesis-at-retrieval model that we will consider here is the model of McClelland (1981). McClelland outlined a connectionist model of memory wherein each trace consists of one "instance" unit representing the trace as a whole along with "property" units for each of the trace's properties. Property units within a trace reinforce each others' activation through bidirectional connections to a central instance unit, while inhibiting the activation of alternative property units from other traces (see Fig. 1). Traces which share a property actually share the same property unit. During retrieval, every trace in memory can therefore become active to the extent that it shares the properties of a recall cue, represented by activating some property units, as well as to the extent that its properties reinforce one another through the instance unit. The inhibition between non-shared properties of different traces also influences the final activation of a trace as similar traces compete for activation.

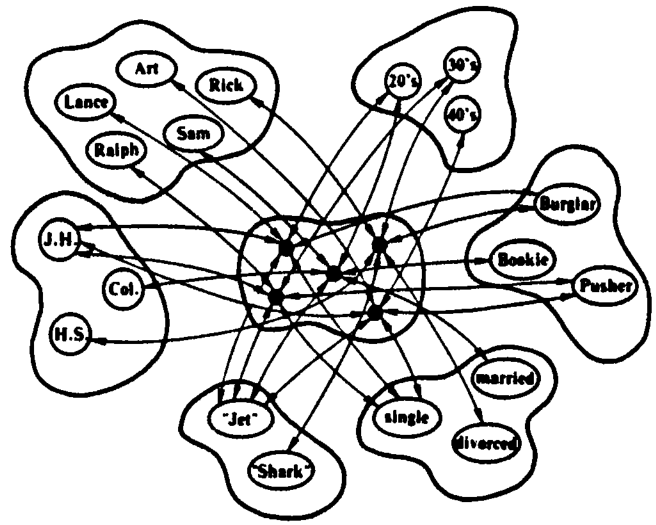

Fig. 1. The synthesis-at-retrieval model of McClelland (1981). The solid black units in the central pool are the instance units, while the units with names are the property units. The units connected with double-headed arrows are mutually excitatory. All the units within the same pool are mutually inhibitory.

Those property units which ultimately remain active after a settling period of activation and competition constitute the recalled information, regardless of whether they all occurred in the same trace. Therefore, if two traces share properties that are all activated by a recall cue, after a settling period, some of one of the trace's unique properties might be active while some of the other trace's unique properties remained active—a situation leading to a blend error.

The question therefore arises: Is there any evidence that blend errors really do exist? Past experiments have suggested that memories for facts and experiences can indeed combine with each other. For instance, the body of literature dealing with story recall, dating back to the work of Bartlett (1932), has demonstrated that subjects tend to misremember portions of stories in accord with their own background knowledge. However, it is not entirely clear to what degree subjects in these experiments had actively elaborated on a story during memory encoding, trying to make sense of an unfamiliar text by integrating their past knowledge with their memory for the story. Similar reservations apply when evaluating more recent work on integrative memory for sentences following Bransford and Franks (1971); the blending of similar sentences in these experiments may well have occurred during encoding. As we noted earlier, though, all models allow for trace synthesis during encoding. Even single-trace models can easily account for "blends" in circumstances where subjects combined information during encoding, because the traces formed during such encoding will themselves already represent "blended" information. For instance, if a subject has consciously combined two sentences sharing an agent, thereby forming a new larger trace, subsequent cueing with that agent can bring forth the combined sentence trace as a *single* trace. Therefore, in order to distinguish single-trace from synthesis-at-retrieval models, the focus of an investigation must be directed to blending of similar separate memory traces at the time of retrieval and not during encoding.

Experiments by Loftus on eyewitness testimony (Loftus, 1977; Loftus, Miller, & Burns, 1978) first introduced the term "blending." These studies suggested that subjects blend their memories of an event with later misleading information when asked to recall details of the event. Loftus claimed that subjects' original episodic memories had been integrated with the postevent information. However, critiques by McCloskey and Zaragoza (1985a, 1985b) raise the valid concern that Loftus' experimental design inadequately controlled a demand characteristic that would lead subjects to respond with blended information simply because they trusted the veracity of the misleading postevent information. (It is worth noting that Metcalfe, 1990, has applied her synthesis-in-storage CHARM model to explain the supposedly conflicting data from both Loftus and McCloskey and Zaragoza.) In any case, Loftus' experiments allowed and may have even encouraged subjects to integrate memories at the time of encoding the postevent information rather than during recall, thereby reducing their relevance to the present discussion.

A recall experiment performed by Anderson and Bower (1971) suggests an interesting test of the models. In this experiment, subjects learned a list of sentences, including some sentences that shared a verb, and were later asked to recall the sentences, given either an agent or an agent plus a verb as the cue. Among their results, Anderson and Bower reported "object intrusion errors": errors when subjects, cued with an agent, correctly recalled the verb from the sentence containing that agent but recalled an object from a different sentence. They found that these intrusions were more common between sentences sharing a verb than between sentences sharing no words. Although this isolated result can be explained in many ways, it seems to naturally follow from the expectations of a synthesis-at-retrieval model, because the intruding object came from a noncued sentence trace that was simultaneously activated as it shared a verb with the cued trace. By contrast, Anderson and Bower (1971) were surprised by the intrusions, although they generated a plausible post hoc explanation for them. Unfortunately, the subjects in this experiment were told that they were participating in a memory experiment before initially studying the sentence lists and, once again, may have consciously integrated similar sentences during encoding. As a result, Anderson and Bower's results cannot be taken as strong evidence for blending during recall.

The present experiments sought to find evidence for or against the existence of blend errors using a recall task like that of Anderson and Bower (1971), while attempting to minimize the possibility that subjects might deliberately integrate sentences during encoding. Imagine the situation wherein two sentences share three out of five content words, and the three shared words all are used as a recall cue. This cue would match the two overlapping sentences equally well. If a subject were asked to recall the remaining two content words from one of the sentences, they could correctly answer with the remaining words from *either* overlapping sentence. For instance:

Sentence 1: The doctor gave the plumber the coat in the lobby.
Sentence 2: The doctor gave the plumber the watch in the kitchen.
Cue: The doctor gave the plumber the _____ in the _____.

Assuming that each sentence is encoded as a single trace, with its constituent words as properties, a synthesis-at-retrieval model of recall would predict difficulty for subjects' recall, because the ambiguous cue would equally activate the traces for both of the sentences and their constituent properties. As a consequence, subjects may be prone to making a "crossover" intrusion error by responding *coat* + *kitchen*, or *watch* + *lobby*, mixing words from both sentences. By contrast, a single-trace model would predict that either of the two traces matching the ambiguous cue would be accessed in a probabilistic search, yet the words from only one of these two traces would be retrieved. In other words, a single-trace model would predict no blending of these highly similar memory traces even with such ambiguous cues.

Following the logic of this example, four experiments were designed to create a situation that would maximally facilitate blending according to a synthesis-at-retrieval model. Experiment 1, which can serve as a prototype for all four experiments, used a list of sentences in which half of the sentences were constructed in pairs sharing three content words with one other sentence, while the other half served as completely dissimilar controls. Unlike the procedure in many earlier memory experiments, subjects learned this list under the pretense of a rating task, unaware that they would need to later recall any sentences. The need to recall sentences was only introduced after a further paragraph rating task used to create a delay between study and test. Subjects therefore had no reason to continually rehearse sentences or to actively integrate the traces of overlapping sentences, thereby allowing the experiment to plausibly investigate blending during retrieval alone. Indeed, early pilot testing confirmed that if subjects knew that they would later need to recall sentences, they often attempted to combine sentences sharing content words into larger composite, "preblended" sentences. Active blending was further discouraged by separating the presentation of overlapping sentences as far as possible. As a check on the use of explicit trace synthesis at encoding, in Experiments 2 through 4, posttest questionnaires were given to assess the degree to which subjects were aware of our manipulations and whether or not they consciously integrated similar sentences. Obviously such questionnaires cannot eliminate all versions of synthesis-at-encoding but they do shed some light on explicit and deliberate strategies.

In all four of the experiments, two general results were of primary interest, each testing an hypothesis of synthesis-at-retrieval models regarding memory blends: (1) Similar memory traces should sometimes be blended; and (2) There should be interference between similar sentences even after accessing one of them. The first hypothesis, asserting the existence of memory blends, would be supported by finding more crossover intrusion errors between the similar overlapping sentences than the base rate of intrusions between dissimilar control sentences. If no difference was discovered, the null hypothesis of single-trace models would be supported instead. The second hypothesis concerned postaccess interference between similar traces: it was expected that subjects should be less likely with overlapping than with control sentences to correctly recall a complete sentence after having accessed its trace. This hypothesis developed in contrast to the claim made by single-trace models that once a trace is accessed, retrieval should simply involve a readout of the relevant information, without interference from any other traces. In a synthesis-at-retrieval model, however, the overlapping sentence traces should continually influence each other during retrieval of the trace properties, whereas the control sentence traces (which shared no properties with other traces) would be relatively immune from such postaccess interference. The second hypothesis was tested by examining the conditional probability that a subject would correctly recall all target words, given that they had correctly recalled one of the words. Synthesis-at-retrieval models, but not single-trace models, predict that this conditional probability should be lower with overlapping sentences than with control sentences.

This paper will present four experiments and then discuss computer simulations of a synthesis-at-retrieval model of the results of these experiments. The first experiment will be presented in more detail than the others, as it serves as a prototype for all four experiments.

Experiment 1

Method

Subjects. Thirty-five undergraduates from Carnegie–Mellon University voluntarily participated as subjects. All subjects were native speakers of English. Subjects were reimbursed for their participation with their choice of either $5 or credit to fulfill an introductory psychology course requirement.

Materials. Thirty-two experimental sentences were randomly created for each subject. (Preliminary testing had determined that 20 sentences could be remembered too easily and allowed conscious integration of overlapping sentences and that 40 sentences were too many to learn and later remember.) Four different sentence templates were used to generate stimuli with moderately differing semantics. These templates included five word positions which were filled by random selection (without replacement) from prearranged lists of words of the appropriate semantic type. Each list of word fillers included only as many words as would be needed to create the 32 sentences, so that each subject's random sentences were generated from the same finite set of words. The four sentence templates (with semantic-type word positions highlighted) included:

(a) The ***PERSON TRANSFER***(-ed) the ***PERSON*** the ***OBJECT*** in the ***LOCATION***.
(b) The ***PERSON CONTACT***(-ed) the ***PERSON*** on the ***BODYPART*** in the ***LOCATION***.
(c) ***PERSON*** helped the ***PERSON VERB*** the ***OBJECT*** in the ***LOCATION***.
(d) The ***PERSON*** saw the ***PERSON VERB*** the ***OBJECT*** in the ***LOCATION***.

The templates were chosen so as to provide a semantically-neutral setting for word fillers. These words were chosen to be maximally dissimilar within a semantic type, while not being infrequent in everyday use. Care was also taken to avoid closely related words in the different semantic types by disallowing the use, for instance, of both "barber" and "barbershop."

Eight sentences were generated using each template, grouped into two sets of four. Two sentences from each group of four (or "quadruple") were selected to become overlap sentences, the other two left as matched control sentences. An intrusion error between two control sentences would only be scored as a crossover error if it occurred between control sentences previously paired in one of these quadruples. In this way, the base likelihood for crossover errors for overlap and control sentences could be most closely matched. Three of the word positions in each quadruple were randomly chosen for the overlap; any three of the five were equally likely to be chosen for overlap (although for convenience, the examples in this paper all show the first three words as overlapping). The three words in the overlapping positions of one of the two overlap sentences were then replaced with the corresponding words from the other sentence, thereby creating a pair of sentences which differed by only two words.

The order of presentation of the 32 sentences was randomized for each subject with the constraint that a minimum of 12 sentences intervened between any two matched overlap or matched control sentences from the same quadruple. Two additional sentences were constructed for use with all subjects as the first and the last study sentence. These two sentences closely resembled the style of the randomly generated sentences, while using words and a template not included elsewhere.

Sentence 0: The astronomer told the usher to shake the keys in the shed.

Sentence 33: The zookeeper told the programmer to lift the rock in the shop.

These sentences served as "buffers," reducing any serial or temporal position effects that might affect the memory salience of the first or last sentence read.

For the second phase of the experiment, a paragraph from a discussion of "art and reality" was selected as a delay stimulus. The paragraph was chosen for its difficult concepts and wording and because its topic seemed closely related to the experiment's cover story.

For the third phase, recall cues (test sentences) contained all words from a previously presented sentence except the two words in the nonoverlapping word positions, which were replaced by blank underlines of a fixed length. Presenting any one sentence from an overlapping pair in this manner therefore acted as a cue for *either* of the two paired sentences. One overlap and one control sentence from each of the eight quads were randomly chosen to become test sentences. The first test sentence shown to subjects always cued Sentence 0, the first buffer sentence. The 16 actual test sentences followed, presented in a completely random order.

All materials were presented on the screen of a standard Apple Macintosh II monitor. All responses were made using the Apple Macintosh keyboard.

Design and procedure. The experiment used a within-subject design. The random generation of sentences created many unlikely combinations of sentence constituents, enabling the experiment to use a convincing cover story. Subjects were told that the experiment was designed to explore how people read sentences that "sound strange," those in which some or all of the content words do not seem to belong together "in reality." Subjects were therefore unaware that they would need to remember any of the sentences later. They were also assured that the experiment did not measure any reaction times, so that they could concentrate on the tasks without worrying about how long they spent working on them.

In the first phase, subjects saw 34 sentences, one at a time. For each sentence, subjects made a judgment about the overall "plausibility" of the sentence, followed by judgments of how appropriate each of the five main words seemed in the context of the rest of the sentence. First the sentence appeared along with a prompt to rate it as a whole, then each of the five words appeared sequentially underneath the sentence along with a prompt to rate it as well. All ratings fell on a five-point scale; subjects were shown sample sentences with example ratings and had to demonstrate their understanding of the scale to the experimenter. The purpose of these ratings was both to make the cover story more convincing and to ensure that the subjects paid attention to every sentence and its primary words, forming memory traces after processing them at a "deep," semantic level. The actual judgment responses were never recorded.

In the second phase, subjects were told to read a paragraph and then answer some questions about it. They read the delay task paragraph then answered seven 5-point judgment questions similar to those in the first phase, evaluating the overall understandability of the paragraph along with that of several of its more obscure words. Due to the difficulty of the selected passage, this task provided a delay period of approximately 5 min during which there was no reason for subjects to actively attempt to remember any of the sentences from the first phase.

In the third phase, subjects were told that they were going to have to remember many of the sentences from phase one. They were presented with 17 test cues one at a time and had to type in the words previously seen in the blank positions in a left-to-right order. Subjects were told that they may have seen two sentences that fit a cue equally well, but that in those cases, they would have to remember both of them, one at a time. However, they were never told that a particular test cue matched two sentences until they had first already recalled one sentence. It was stressed that in choosing each response they should take care to recall two words from within the same sentence in the first phase. So after the subject first recalled two words from one overlap sentence, they would be told that they had seen another sentence that also matched the cue and be asked to now recall two more words belonging to the other matching sentence. In this manner, all of the overlap sentences from the first phase were tested, along with one half (randomly selected) of the control sentences. The experiment therefore included three response conditions: (1) the first pair recalled given an overlap cue (the *first overlap* condition); (2) the second pair recalled given an overlap cue (the *second overlap* condition); and (3) the pair recalled given a control cue (the *control condition*).

After typing in each word, subjects were asked to rate their confidence that they had correctly recalled the word, again on a five-point scale. This was done to encourage subjects to carefully consider each recall; as expected, confidence was higher for words correctly recalled than for words recalled incorrectly. After typing in each set of two words, they were also asked to rate their confidence that they had correctly recalled both words *from the same single sentence in the first phase*. This served both as a reminder to subjects that we wanted them to remember a whole sentence, not mixtures of sentences, and as a dependent measure for later analysis.

Finally, after all the sentences were tested, subjects were given a questionnaire which attempted to discover to what extent they were aware of the overlapping sentence manipulation during the first phase of the experiment. Questions asked whether they became aware that more than one sentence used the same words and whether or not they tried to remember or integrate earlier sentences sharing words with later overlapping sentences.

Results

Overall, subjects correctly recalled individual words 52% of the time. They correctly recalled *both* target words (thus correctly completing a sentence) 34% of the time, and correctly recalled neither word 30% of the time. For the remaining 36% of the time, subjects recalled one word correctly and either left the other word blank, guessed a word that had never appeared in the study sentences, or responded with a word from another sentence. This last category included any crossover errors.

The overall error rates were not the same in each experimental condition, however. Subjects were able to recall more sentences correctly in the first overlap condition (M = 42%) than in the second overlap condition (M = 19%). They performed about the same in the control condition (M = 41%) as in the first overlap condition. A Wilcoxon signed-ranks test of these differences across subjects showed that the difference between success rates in the first and second overlap conditions was significant (z = 4.261, $p < .0001$), as was the difference between success rates in the second overlap and control conditions (z = 4.261, $p < .0001$), but the difference between first overlap and control conditions' success rates was not significant (z = .027, $p > .05$).

The test of the first hypothesis distinguishing synthesis-at-retrieval from single-trace models compared the rates of crossover errors relative to these conditions. Subjects made crossover errors more often when tested in the overlap conditions (5.4% with the first, 2.1% with the second) than in the control condition (0.7%). The raw frequency of crossover errors was rather small (15, 6, and 2, respectively), but the difference between the rates in the first overlap versus control conditions was significant nonetheless (using the sign test across subjects, $p < .01$). On the other hand, the difference between the rates in the second overlap versus control conditions was not significant ($p > .05$). The frequency of crossover errors was also significantly greater in the first overlap than in the second overlap conditions ($p < .05$).

In order to show that the greater frequency of intrusions between overlap sentences was not simply a result of a relative tendency to recall words from overlap sentences, we examined the frequency of another type of intrusion within a matched set of four sentences: a word from an overlap sentence intruding into recall of a control sentence, or vice versa. In fact, words from overlap sentences intruded into recall of a matched control sentence only 1.4% of the time, while words from control sentences intruded into recall of a matched overlap sentence 2.7% of the time.

The second hypotheses test compared, across conditions, the conditional probability of correctly recalling two words from a target sentence, given the correct recall of one word. For overlap sentences, this probability was higher in the first pair of words recalled than in the second pair (p = .50 in the first overlap condition; p = .36 in the second overlap condition); with control sentences, the probability was .58. To test the significance of these differences, the conditional probability was calculated for each condition and each subject separately, and Wilcoxon signed-ranks tests were performed across subjects. The conditional probability in the control condition was significantly higher than in either the first overlap (z = 1.81, $p < .05$) or the second overlap conditions (z = 4.11, $p < .0001$); the conditional probability in the first overlap condition was significantly higher than in the second overlap condition (z = 3.10, $p < .001$).

Analysis of confidence ratings focused on subjects' confidence that both recalled words came from the same sentence. In general, the ratings corresponded quite well to the actual correctness of their recall. An examination of the first overlap condition confidence ratings shows that the average confidence for a answer containing both correct words (M = 4.5, SD = .87) was higher than the average confidence for an answer containing no correct words (M = 2.3, SD = 1.4). Subjects were less confident in crossover error responses (M = 3.7, SD = 1.3) than in correct responses, though still more confident in these crossover errors than in other answers where one word was wrong (M = 3.3, SD = 1.3). In fact, six of the 15 blend errors were made with a full confidence of 5. Crossover errors in the other conditions were too infrequent for useful analysis.

Discussion

The results of this experiment confirmed the hypotheses of synthesis-at-retrieval, not single-trace, models. First of all, subjects made significantly more crossover errors with the similar, overlap sentences than with the dissimilar control sentences. The intrusion of one incorrect word into an overlap sentence recall was also more likely to come from the paired overlap sentence than from any other single sentence in the study list. The greater number of intrusions between overlap sentences than control sentences did not result from a general tendency to recall words from overlap sentences. All of this can be explained by the prediction of synthesis-at-retrieval models that similar traces will blend during recall because every trace contributes to the recalled information to the extent that its properties are similar to the cue and to those of other traces. Single-trace models would not expect any differences in the intrusion rates based on similarity of the traces, because they preclude any interactions between traces once a single trace is selected for access.

Second, the conditional probability that subjects remembered both words after correctly recalling one word was higher with control sentences than with the overlap sentences. This result contradicts the claim of single-trace models that a trace is retrieved as a whole; instead it fits the prediction of synthesis-at-retrieval models claiming that the similarity between overlap sentences should cause an increased chance of retrieving words from different sentences. The difference between the first recall attempt with overlap sentences and recall with control sentences was not very large, however.

Perhaps a better measure of performance with overlap sentences might be the *average* probabilities of both first and second recall attempts, a measurement that would enhance the difference between results with overlap versus control sentences. In fact, it is extremely difficult to know which measure, either of first recall alone or of the average of first and second recalls, would be less biased. With overlap sentence cues, subjects were allowed to recall either of two sentences, while control cues must have been used to access a single sentence. Clearly, subjects correctly recalled words more frequently in the first overlap condition than in the second overlap condition because the first recall reflects retrieval of the stronger of two traces. Therefore if one just looks at the performance of subjects in the first overlap condition, one potentially biases the analysis towards an examination of retrieval of only strong traces. On the other hand, it is not clear that averaging performance across qualitatively different recall attempts can provide valid comparisons between the overlap conditions and the control condition. Furthermore, examination of the first attempt alone allows comparison of crossover intrusion rates between two conditions with the same overall level of recall accuracy.

The analysis of confidence ratings indicates that subjects demonstrated a general ability to judge the correctness of their answers. Focusing on just the first overlap condition, we see that subjects were less confident with blend errors than with correct recalls. This raises the concern that subjects might have been aware of their incorrect answers when making crossover errors, consciously deciding to answer with a word from a matched overlap sentence just because it seemed a more appropriate guess than any other wrong answer. However, subjects were more confident on average in crossover errors than in other one-word errors, and some crossover errors were given with full confidence. In synthesis-at-retrieval models like that of McClelland (1981), a subject's confidence might be most naturally captured as a function of the overall "goodness" (cf. Rumelhart et al., 1986) of the state of the memory system. In such a model, the goodness associated with a blend error will generally be higher than that of any other error, but cannot be as high as with a correct answer due to the inhibition between the simultaneously active sentence traces forming the blend. The pattern of confidence ratings as obtained here cannot be used to distinguish between the models of interest and, since the same pattern held in each of the next three experiments, we will not discuss them further.

Experiments 2–4

The following three experiments used variations on the materials or procedure used in Experiment 1. Experiment 2 was designed to allow only one correct response in both the overlap and the control conditions to eliminate the difficulty in knowing how to interpret data from the two different overlap conditions. Experiment 3 attempted to increase the rate of blending by reducing potential response competition between the target words in overlap sentence traces. Experiment 4 eliminated the rating of individual words in the first phase to test the possibility that subjects may have been encoding single words as traces, as opposed to whole sentences. The methodology of each will be described below only in so far as it differed from Experiment 1.

Experiment 2

In order to allow only one correct response in the overlap condition, as well as the control condition, unambiguous overlap sentence test cues were used, in which the number of shared content words was reduced from three to two. For example, an overlap pair would now look like this:

Sentence 1: The doctor gave the plumber the coat in the lobby.

Sentence 2: The doctor gave the lawyer the watch in the kitchen.

while the cue remained:

The doctor gave the plumber the _____ in the _____ .

Note that this cue only matches the first overlap sentence completely, while offering a partial match to the second sentence. We hoped that the partial match would be sufficient to induce blend errors, even while it eliminated the availability of two alternative correct responses.

The number of sentences was increased to 36 to partially compensate for the decrease in the amount of data generated now that each overlap sentence cue allowed only one correct answer. In addition, one of the overlap sentences in each overlap pair was used as a mate for a control sentence, since it could not share any content words with the control sentence anyway. This allowed a significant reduction in the number of control sentences, half of which were never tested in Experiment 1 and would have yielded no data anyway.

Six different sentence templates were used. These included the four used in Experiment 1, plus two new ones:

(e) The *PERSON VERB*(-ed) the *OBJECT* in the *LOCATION* with the *PERSON*.
(f) The *PERSON VERB*(-ed) the *OBJECT* in the *LOCATION* for the *PERSON*.

Six sentences were generated using each template, grouped into two sets of three. Two sentences from each group of three were selected to become overlap sentences; the other was left as a matched control sentence.

The order of presentation of the 36 sentences was randomized for each subject in the following fashion: 12 control sentences and 12 overlap sentences (one "overlap test" sentence from each overlap sentence pair) were presented as the first 24 sentences, followed by the rest of the overlap sentences (the "overlap distractors"). This allowed the overlap distractor sentences to provide retroactive inhibition, both for the overlap test sentences and the control sentences. The order was then randomized with the further constraint that a minimum of 12 sentences intervened between overlap test or control sentences and their paired overlap distractor sentence. Four buffer sentences were constructed for use as the first two and the last two study sentences, as with the two buffer sentences in Experiment 1.

In addition, a posttest questionnaire was added at the end of the experiment, to assess how aware of our manipulations subjects had been. It included eight questions, asking subjects whether they were aware of sentence overlapping, asking them to give their estimates of the nature and frequency of the overlaps, and asking whether evaluating a later overlap sentence caused them to recall or consciously integrate an earlier overlap sentence (see Appendix A).

Thirty-eight subjects from the same source as Experiment 1 were used.

Results

The subjects performed worse overall than those in Experiment 1: they correctly recalled individual words 42% of the time. They correctly recalled both target words 25% of the time and neither word 42% of the time. As in Experiment 1, error rates differed between conditions. With control sentences, subjects correctly recalled both words 29% of the time, significantly more ($z = -2.421$, $p < .01$) than the 22% correct recall with overlap test sentences. Fully 9% of subjects' answers with overlap target cues consisted of recalling *both* words from its paired overlap distractor sentence, despite the fact that the cue unambiguously matched the overlap target sentence. This occurred only once (0.2%) with control sentences.

The first hypothesis test measured the frequency with which a word from an overlap distractor sentence was recalled with a correct word, given either an overlap test cue or a control cue. These crossover errors were more frequent between matched overlap sentences (18 occurrences, or 8.0%) than between overlap distractors and their matched controls (6 occurrences, 1.3%). As in Experiment 1, the raw frequency of blend errors was small, but the difference between the crossover error rates was significant (sign test, $p < .05$). This result again supported synthesis-at-retrieval models, in that subjects made intrusion errors more often with (similar) overlap than with (dissimilar) control sentences.

The conditional probability of correct recall of two, given one, words was essentially the same in both conditions: .43 with overlap test sentences, and .44 with control sentences. The difference between these probabilities was not significant ($z = 1.07, p > .05$). This matched the prediction of single-trace models that postaccess interference should be unaffected by shared properties.

The posttest questionnaires indicated that most subjects were unaware of the precise nature of sentence overlapping during the first phase of the experiment. No single subject got all of the questions correct. Of the 34 subjects for which we had complete data, all but one noticed that at least some sentences shared content words. However, their estimates of how many sentences shared words ranged from 5 to 35; only seven subjects correctly estimated that overlap sentences shared words with one and not more than one other sentence; and all but four subjects thought that different overlap sentences shared different numbers of words. Nine subjects reported that presentation of a second overlap sentence caused them to remember the earlier overlap sentence, yet only four of these reported that they sometimes tried to integrate the two sentences into a larger whole. The nine subjects who reported recalling earlier overlap sentences were no more likely to make blend errors than other subjects ($\chi^2(1) = .58, p > .05$); eliminating them from earlier data analyses had no significant impact on the results of the hypothesis tests. Similarly, subjects who reported that they had consciously integrated overlap sentences were also no more likely to make a blend error than other subjects ($\chi^2(1) = .81, p > .05$), and eliminating them from the analyses had no significant impact on the results. The posttest questionnaire therefore indicated that the cover story and other manipulations that were designed to minimize blending during encoding were successful with almost all subjects, at least at a level accessible by conscious reporting. The same general pattern of responses to the questionnaire held across both of the following experiments.

Experiment 3

The number of blend errors obtained in the first two experiments was admittedly less frequent than we had originally expected to observe, considering that the experimental design had been constructed to maximally facilitate blending during recall. However, both experiments involved a situation in which the constituent words of two overlap sentences directly competed for recall selection, a competition that may have placed a severe limitation on the possibility of blending similar traces. According to many synthesis-at-retrieval models, if one sentence's trace is initially more active than the other, the mutual inhibition between the sentences can overamplify this initial advantage—a "rich-get-richer effect" (Grossberg, 1976; McClelland & Rumelhart, 1988). One would therefore expect fewer blend errors, as the stronger trace's properties would suppress the activation of competing properties.

Experiment 3 tried to avoid this potential intersentence competition by using overlap sentences with nonoverlapping portions that corresponded to different semantic types. The 32 sentences used four different sentence templates, each including three mandatory content words and zero, one, or two optional content words. The four sentence templates included:

(a) The *PERSON VERB*(-ed) the *OBJECT*.
(b) The *PERSON VERB*(-ed) the *OBJECT* in the *LOCATION* with the *PERSON*.
(c) The *PERSON VERB*(-ed) the *OBJECT* in the *LOCATION*.
(d) The *PERSON VERB*(-ed) the *OBJECT* with the *PERSON*.

These templates were further grouped into two pairs: (a) 3-word with (b) 5-word, or (c) 4-word/location with (d) 4-word/accompaniment, so that each of the optional arguments only appeared in one member of a matched pair. These pairs were then grouped into sets of four sentences as in Experiment 1, with two sentences (one from each template) selected for overlap and two left as a matching control. In the overlap sentences, only two of the first three word positions were randomly chosen for overlapping, so that overlap sentences differed by at least one mandatory word in order to be able to test the 3-word sentences from template (a). As an example, using the (c) + (d) template, we could generate the following overlap pair:

Sentence 1: The doctor moved the coat in the lobby.
Sentence 2: The doctor moved the watch with the plumber.

Test cues contained all words from a previously presented sentence except those in the three nonoverlapping word positions. The optional arguments were indicated to subjects by placing parentheses around the function words in the phrase. For example, the cue:

The doctor moved the _____ (in the _____) (with the _____).

could elicit either of the paired overlap sentences as a correct response, just like in Experiment 1. Because subjects had to be able to explicitly indicate the absence of an optional argument to correctly recall sentences from most of the templates, when they left a blank response they had to state either that they were unable to recall a word or that they actually believed that there had been no word in that position in the original sentence.

Except for one of the mandatory content words (the object, in the example above), the nonoverlapping arguments were not in conflict with one another, theoretically allowing both to be simultaneously active regardless of the inhibition between potential fillers. Therefore, subjects could have occasionally recalled arguments from both study sentences (such as "the coat in the lobby with the plumber"), even if instructed to recall words from only one sentence. Subjects might also have made blend errors resembling those found in the previous experiments ("the coat with the plumber" or "the watch in the lobby").

Twenty-nine subjects from the same source as the other experiments were used.

Results

Overall, the subjects performed about the same as subjects in Experiment 1. They correctly recalled each individual word 50% of the time. They correctly recalled all three target words 31% of the time and correctly recalled none of the words 45% of the time. In the control condition, subjects were able to correctly recall all three words 38% of the time, while in the first overlap condition they correctly recalled 42% and in the second overlap, only 13%. The difference between the rates in the control and first overlap conditions was not significant ($z = .81, p > .05$), but the differences between the control and second overlap ($z = 4.20, p < .0001$) and between the first and second overlaps ($z = 4.29, p < .0001$) were quite significant.

To determine the rates of blending in this experiment, it was first necessary to decide what types of errors would be considered crossover errors. It was decided that any response mixing three words (or blanks, in the case of optional arguments) from two matched sentences would be scored as a crossover error. For a complete list of the various types of crossover errors, and their obtained frequencies, see Table 1.

Crossover errors were more frequent (sign test, $p < .001$) in the first overlap condition (13 occurrences, 5.6%) than in the

TABLE 1
CROSSOVER ERROR FREQUENCIES BY TYPE AND CONDITION

Type of error	Condition: 1st overlap	2nd overlap	Control
3 & 5 word sentences			
P V O_1 – –			
P V O_2 L P			
P V O_2 – –	1	1	1
P V O_1 L P	0	0	0
P V O_1 L –	1	0	0
P V O_1 – P	0	0	0
P V O_2 L –	2	0	0
P V O_2 – P	1	0	0
4 word sentences			
P V O_1 L –			
P V O_2 – P			
P V O_1 – P	1	0	0
P V O_2 L –	2	0	0
P V O_1 L P	2	0	0
P V O_2 L P	1	0	0
P V O_1 – –	0	1	0
P V O_2 – –	2	0	0
Total frequencies	13	2	1

control condition (one occurrence, 0.4%). There was no significant difference ($p > .05$) between the crossover error rates in the second overlap condition (two occurrences, 0.9%) versus the control condition, while the difference between the rates in the first and second overlap conditions was significant ($p < .01$). Once again, the hypothesis of synthesis-at-retrieval models was supported by the increased number of crossed intrusion errors between similar sentences.

The second hypothesis test compared the conditional probabilities of correct recall of all *three* words, given correct recall of any one of them. This probability did not significantly differ ($z = -.27$, $p > .05$) between the first overlap condition ($p = .47$) and the control condition ($p = .46$). On the other hand, the difference between the conditional probability in the control and second overlap ($p = .22$) conditions was significant ($z = 3.39$, $p < .001$), as was the difference between the first and second overlap conditions ($z = 3.56$, $p < .001$). If one compares just the first overlap with the control condition, single-trace models are supported by the lack of a difference in post-access interference. But as with the first experiment, one could argue that the proper comparison should be between the average of the two overlap conditions and the control condition, in which case the prediction of synthesis-at-retrieval models was supported.

The posttest questionnaires indicated that most subjects did not consciously integrate sentences during encoding. In this experiment, seven subjects reported that presentation of a second overlap sentence caused them to remember the earlier overlap sentence, and eight reported that they sometimes tried to integrate the two sentences into a larger whole. The subjects who reported recalling earlier overlap sentences were no more likely to make blend errors than other subjects ($\chi^2(1) = .62$, $p > .05$ for the first overlap condition; the frequency of errors was too small to analyze in the other conditions); eliminating them from earlier data analyses had no significant impact on the results of the hypothesis tests. Similarly, subjects who reported that they had consciously integrated overlap sentences were also no more likely to make a blend error than other subjects ($\chi^2(1) = .07$, $p > .05$), and eliminating them from the analyses had no significant impact on the results. Thus, although the percentage of subjects who claimed to have consciously integrated overlap sentences was higher in this experiment than in Experiment 2, they were even less likely to make blend errors than those in the earlier experiment.

Experiment 4

In this experiment, we modified the evaluation procedure used in the study phase as a cover story and as insurance of complete encoding. In the other experiments, subjects had been asked to rate the plausibility of a whole sentence and, then, of *each individual content word;* this latter requirement could conceivably have caused subjects to encode a separate individual memory trace for each content word. If this had been the case, both multiple- and single-trace models might explain our results by proposing that subjects were occasionally able to exploit associations between traces for individual words, along with more complete traces for the sentences containing them. Therefore, Experiment 4 duplicated the methods of Experiment 1, except that it eliminated the rating of individual words during the study phase. Thirty-seven subjects were used, from the same source as the other experiments.

Results

Overall, the subjects performed worse than the subjects in the other three experiments. Subjects correctly recalled each individual word only 32% of the time. The lower performance can be attributed to the fact that subjects, because they no longer had to evaluate individual words, spent less time encoding the sentences in the first phase. They correctly recalled both target words 16% of the time and correctly recalled neither word 52% of the time. In the control condition subjects were able to recall both words correctly 17% of the time, in the first overlap condition, 26% of the time, and in the second overlap condition, only 5% of the time. All pairwise comparisons of these success rates were significant: control vs. first overlap ($z = -2.785$, $p < .005$), control vs. second overlap ($z = 3.81$, $p < .0001$), and first vs. second overlap ($z = 4.395$, $p < .0001$).

Even though subjects made more errors overall, there were fewer blend errors in this experiment. No crossover errors were obtained between control sentences. Consequently, crossover errors were significantly more frequent ($p < .01$) in the first overlap condition (7 occurrences, 2.4%) than in the control condition. They were also more frequent ($p < .05$) in the second overlap condition (5 occurrences, 1.7%) than in the control condition. There was no significant difference ($p > .05$) between the crossover error rates in first and second overlap conditions. Therefore, this experiment eliminates the concern that crossover errors might have been an artifact of the specific instruction to rate words individually in Experiments 1 through 3 because crossover errors, as predicted by synthesis-at-retrieval models, were obtained even though subjects were instructed only to evaluate the sentences as wholes.

The conditional probabilities of correct recall of both words, given correct recall of one word, did not significantly differ ($z = -1.16$, $p > .05$) between the first overlap condition ($p = .35$) and the control condition ($p = .31$). On the other hand, the difference between the conditional probability in the control and second overlap ($p = .16$) conditions was significant ($z = 2.78$, $p < .005$), as was the difference between the first and second overlap conditions ($z = 2.93$, $p < .005$). As with Experiment 3, support for single-trace versus synthesis-at-retrieval models would therefore depend on whether one compared the first overlap only, or the average of the first- and second-overlap conditions to the control condition.

The posttest questionnaires indicated that once again, most subjects did not consciously integrate sentences during encoding. All but one subject noticed that at least some sentences shared content words, yet their estimates of the details of the overlapping was again widely inaccurate. In this experiment, eight subjects reported that presentation of a second overlap sentence caused them to remember the earlier overlap sentence, and only four reported that they sometimes tried to integrate the two sentences into a larger whole. The eight subjects who reported recalling earlier overlap sentences were no more likely to make blend errors than other subjects ($\chi^2(1) = .25$, $p > .05$, for the first overlap condition); eliminating them from earlier data analyses had no significant impact on the results of the hypothesis tests. Similarly, subjects who reported that they had consciously integrated overlap sentences were also no more likely to make a blend error than other subjects ($\chi^2(1) = .11, p > .05$), and eliminating them from the analyses had no significant impact on the results.

General Discussion

Taken together, the four experiments reported here present an initially puzzling pattern of results. Table 2 illustrates the pattern across all four experiments for the major dependent measures discussed here. The rates of crossover errors obtained in all four experiments supported the hypothesis of synthesis-at-retrieval models that similar sentence traces should blend during recall. We obtained the predicted blend errors, but not a very large number of them—only about one out of 20 responses in the first overlap conditions.

The pattern of conditional probabilities indicating the presence or absence of postaccess interference between sentences showed less consistency across experiments and, therefore, only inconclusive evidence for the effect predicted by synthesis-at-retrieval models. If one wishes to compare the average of the first and second overlap conditions with the control condition, one will find support for synthesis-at-retrieval models in Experiments 1, 3, and 4, because postaccess interference was found with the average overlap sentences. If one instead compares only the first overlap condition with the control condition, support for synthesis-at-retrieval models can only be found in Experiment 1.

One must consider the possibility that the obtained pattern of results was somehow a consequence of encoding blends, of the sort that we had tried to avoid in designing the cover task. Yet the results of posttest questionnaires showed very little indication of any synthesis of traces during encoding. Subjects generally did not, in fact, recall and reintegrate earlier overlap sentences while reading later ones—and the obtained results were unaffected by excluding those subjects who did report occasional blending during encoding. So we may tentatively conclude that the pattern of responses and the frequency of crossover errors was obtained as a result of "true" blending, either in storage or during retrieval.

Given this conclusion, what implications do the data have for the various models of recall? We will first examine in some detail how a synthesis-at-retrieval model might account for the data. Later, we will turn to a discussion of whether single-trace and synthesis-in-storage models might produce similar results.

Computer Simulations of a Synthesis-at-Retrieval Model

In order to determine whether or not a detailed implementation of a synthesis-at-retrieval model could account for the data, we constructed an interactive activation model based on the synthesis-at-retrieval model discussed in the introduction (McClelland, 1981). The resulting computer simulations proved to be very informative as we attempted to make sense of the experimental data.

TABLE 2
Summary of Results for Experiments 1–4

	Experiment number and condition[a]										
	1			2		3			4		
Dependent measure	C	1	2	C	1	C	1	2	C	1	2
Percentage correct (word-by-word)	41	42	19	29	22	38	42	13	17	26	5
Percentage crossover intrusion errors	0.7	5.4	2.1	1.3	4.0	0.4	5.6	0.9	0.0	2.4	1.7
Conditional probability (2 correct | 1 correct)	.58	.50	.36	.44	.43	.46	.47	.22	.31	.35	.16

[a] C = control, 1 = 1st overlap, 2 = 2nd overlap.

The synthesis-at-retrieval model of McClelland (1981) was adapted for the current experiments by assuming that the study of each sentence created an "instance" unit for the sentence trace as a whole, linked with bidirectional excitatory connections to five "property" units, one for each major content word in the sentence. All of the instance units together formed a pool of units, as did all of the property units corresponding to a given sentence role. All units within a pool were linked by bidirectional inhibitory connections (using negative weights), reflecting the fact that the units within them represent mutually exclusive information that should not be recalled simultaneously.

Because the overlap sentences shared words, the corresponding units in the property units pools would also be shared. Simulation of recall proceeded by activating the property units representing the recall cue words. In the case of an overlap sentence, activation from the cued units would be sent along excitatory connections to the instance units for both overlap sentences. For illustration, examine the example of a single quadruple of sentences from Experiment 1 shown in Fig. 2. The two instance units would compete for activation, due to their inhibitory connections; they would also feed back activation to the cue words, as well as the target words. In the pools containing the target words, further competition would result from the activation of the mutually inhibitory units. Eventually,

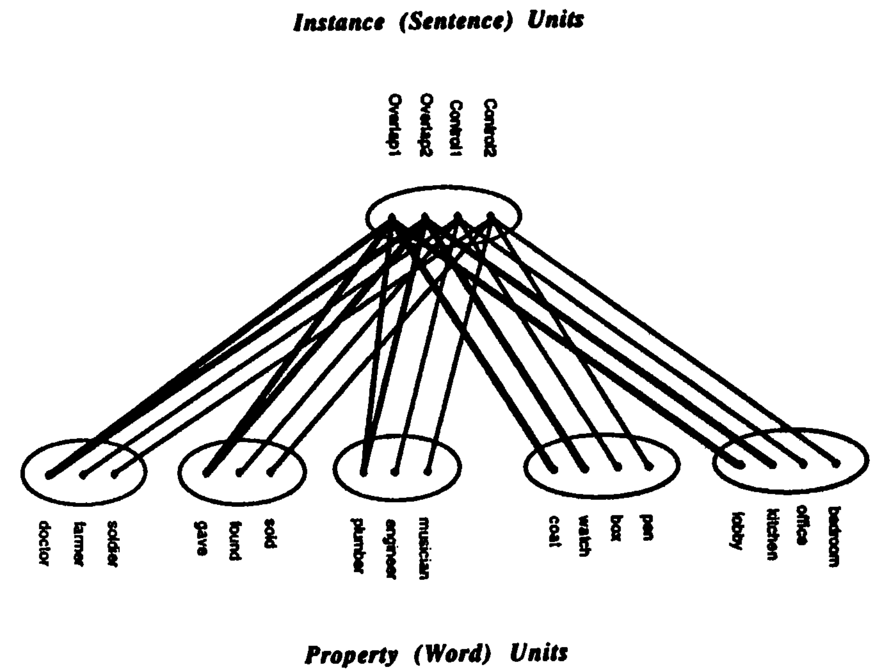

Fig. 2. Interactive activation model of one set of four sentences: two overlap (bold lines) and two control (thin lines). Only excitatory connections are illustrated here; in addition, there are inhibitory connections between every unit within the same pool of units.

the activation levels of all units will stop fluctuating, after the network has "settled," and the units with the highest activations can be chosen as the response. Note that if the cue had been words from a control sentence then the correct answer would be retrieved without competition, because activation would flow to only one instance unit, and only one word per pool.

Actually, as outlined above, the model would not be able to settle the competition between overlap sentence words, because their inputs and resulting activations would always be equal. Based on this, the model led us to expect that our experiments would generate more blend errors than were actually obtained. However, a general problem with the interactive activation model has recently been uncovered in other applications: it fails to take account of the important role of inherent variability in processing (Hinton & Sejnowski, 1986). Once this variability is introduced to the model, several difficulties are resolved (McClelland, 1991).

In the present application, adding variability causes the model to tend to favor one of two complete sentence traces rather than a blend; the blend states represent less-optimal points in the "goodness" landscape of network states (Rumelhart et al., 1986), and variability allows the network to escape such local minima. McClelland (1991) indicates that variability may be introduced in a variety of ways with similar results; in the present model, we simply injected a small amount of normally distributed random noise into the input to each unit at each update. (The detailed assumptions governing the activation process are included as Appendix B.) The addition of intrinsic variability allowed the model to correctly produce lower blend rates, yet it tended to produce too large a difference in the probability of correct recall in the control versus the overlap conditions. (This is equivalent to overpredicting the difference in the conditional probability of recalling both words.) Despite hundreds of simulations using many variations of parameter combinations, we were simply unable to obtain an adequate fit to both the crossover error frequency and the conditional probability data simultaneously.

The inadequacies of the model were eventually traced to the concept of encoding failure. It is quite evident from the experimental results that subjects did not have perfect memory for all sentences in the experiments, considering the rate of 50% or greater recall failure. In any model of memory, an item cannot be recalled if it was improperly encoded in the first place. Yet the models as discussed so far have simply assumed that subjects completely encoded all sentence information. Therefore, we needed to examine how a synthesis-at-retrieval model performs when 50% or more of the properties in traces are either missing or too weak to be recalled correctly.

How would a synthesis-at-retrieval model perform with encoding failures? Because of the simultaneous access of all relevant traces in a synthesis-at-retrieval model, whenever one of the two target words in an overlap sentence had not been adequately encoded, the corresponding word from the other overlap sentence trace would tend to be retrieved instead, as long as it had been properly encoded itself. In fact, this type of effortless default assignment is a well-known advantage of synthesis-at-retrieval models related to their generalization abilities (McClelland, 1981; Rumelhart et al., 1986). However, the synthesis-at-retrieval model discussed earlier was already predicting blend errors when encoding failure was not considered. Would they not predict entirely too many blend errors, relative to the low frequency actually obtained in the experiments, if a good portion of the properties of traces were absent yet filled in through default assignment?

To capture encoding failure in our model, we assumed that each connection between a property unit and its associated instance unit had a certain percentage chance of being absent (20%, one word from each sentence on average). The weights on these connections were simply set to zero. The settings for other weights followed naturally from characterizing the nature of the stimuli in the experiments, although the particular values that we settled on were the result of trial-and-error parameter searching. The weights on intact, properly encoded connections were set to 0.8. The weights on connections within pools of word units were all set to -1.0, while those within the instance unit pool were stronger (-2.1), to counterbalance the activation coming from the three word units and to ensure adequate competition between sentences. In the experimental data, subjects' errors indicated that there were some associations between words from completely different sentences; these associations were captured by inserting randomly determined, nonnegative weights ($M = 0.2$, $SD = 0.25$) between all word and instance units.

The model networks for each experiment reflected the materials used in that experiment. For Experiments 1 and 4, the network therefore contained a pool of 32 instance units, two pools of 32 target word units, and three pools of 24 word units (16 control words + 8 overlap words per pool). For Experiment 2, the network consisted of 36 instance units, two pools of 24 overlap cue units (12 overlap + 12 control), one pool of 36 nonoverlap cue units, and two pools of 36 target units. (Experiment 3 was not modelled due to the difficulty of determining how to simulate the "recall" of absent optional words.) Because the differing results from Experiments 1 and 4 were thought to have been a result of less rigorous encoding in Experiment 4, the only difference between the networks for Experiments 1 and 4 was an increase in the encoding failure rate for Experiment 4 (35% failures, up from 20%).

To test recall, external activation was input to the three cue word units for each of the test sentences. The response of the computer subject was simply taken as the unit with the highest activation above a response threshold (0.1) in each target word pool after 100 time cycles, enough time to allow the network to settle into an equilibrium and form a reasonable response hypothesis. A total of 100 computer subjects were run for each computer simulation; with each computer subject, a new random set of stimulus and random association weights was generated. The performance of each network was fit to the data as shown in Fig. 3 through 5; the dependent measures fit by the model included the frequency of recalling two, one, or zero words correctly, along with crossover intrusion errors (including "double intrusions" of two overlap distractor words in Experiment 2), in both the first overlap and control conditions. In so doing, we ensured that the model not only had the correct overall recall rates but also the appropriate performance with respect to both hypothesis tests—blend rates and postaccess interference rates (which depend on the rates of recalling one or two words correct per sentence). The second

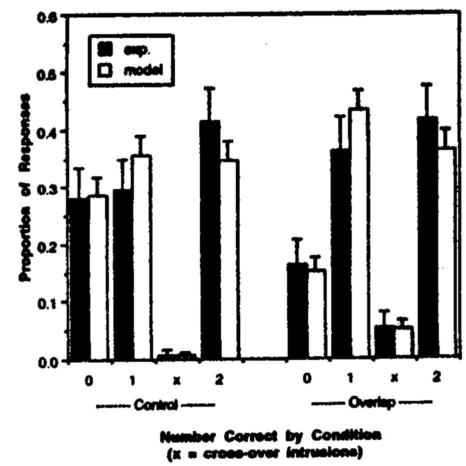

FIG. 3. A comparison of human versus computer subjects for Experiment 1. The proportion of answers sorted by correctness are shown for the first overlap and control conditions. Error bars indicate the 95% confidence intervals.

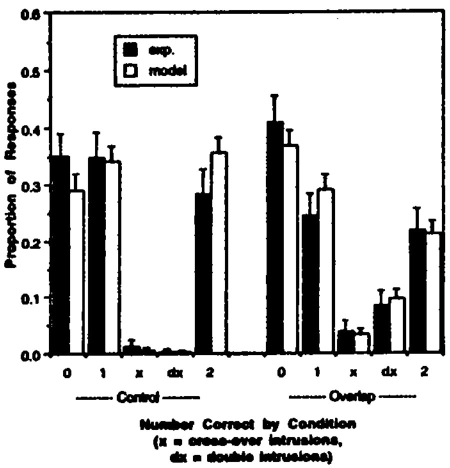

FIG. 4. A comparison of human versus computer subjects for Experiment 2. The proportion of answers sorted by correctness are shown for the first overlap and control conditions. Error bars indicate the 95% confidence intervals.

overlap condition was not modelled; this simplified the simulations, but also provided a more interesting test of the synthesis-at-retrieval model's adequacy; after all,

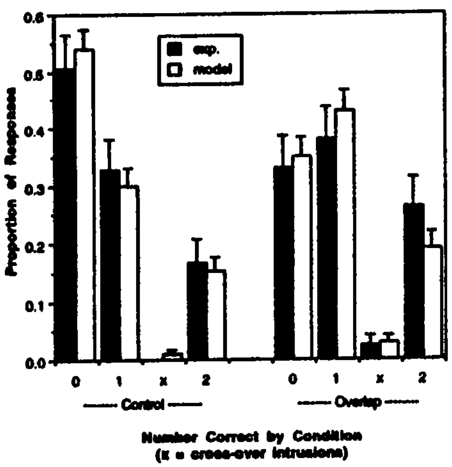

FIG. 5. A comparison of human versus computer subjects for Experiment 4. The proportion of answers sorted by correctness are shown for the first overlap and control conditions. Error bars indicate the 95% confidence intervals.

when we had compared the conditional recall probabilities of only the first overlap to the control conditions, the data had seemed to support single-trace models in three of the four experiments.

In fact, our synthesis-at-retrieval model was able to capture the details of the experimental data quite well. In the data fits illustrated in Figs. 3 to 5, the model's performance was not statistically different from that of our human subjects (for Experiment 1, $\chi^2(7) = 7.0$, $p = .42$; for Experiment 2, $\chi^2(9) = 15.9$, $p = .07$; for Experiment 4, $\chi^2(7) = 11.2$, $p = .13$). However, the fit comes close to failing significance tests in two of the three cases. We were able to obtain even closer data fits by increasing the inhibition between instance units in Experiments 1 and 4 (to -2.5) and decreasing it for Experiment 2 (to -2.0). With the use of this one additional free parameter we were able to achieve a nearly perfect correspondence between the performance of the model and of the experimental subjects (for Experiment 1, $\chi^2(7) = 1.821$, $p = .97$; for Experiment 2, $\chi^2(9) = 2.499$, $p = .98$; for Experiment 4, $\chi^2(7) = 2.996$, $p = .89$).

One reason for varying this inhibition parameter between experiments is that it might reflect the ability of a subject's memory system to strategically increase competitive inhibition between equivalently cued overlap traces, as found in Experiments 1 and 4. While the overlap sentences in Experiments 1 and 4 shared three words, those in Experiment 2 shared only two; the recall cues in Experiments 1 and 4 were ambiguous, while those in Experiment 2 fully cued only one of the overlap sentences. When the recall cues were given, therefore, there was a tougher conflict to be resolved between the sentence traces in Experiments 1 and 4. This leads us to the speculation that the memory systems of subjects may have been attuned to either the total amount of competitive processing or to the ambiguity of cues and were able to adjust the severity of competition accordingly.

Certainly, the suggestion that there is a difference in strategy between Experiments 1 and 4 on the one hand and Experiment 2 on the other is post hoc and would need to be confirmed through followup studies before it could be taken seriously. Further, the utility of the notion that aspects of strategic control might be successfully modelled in terms of regulating the strength of inhibition is only suggested, rather than demonstrated, by the present simulation results. However, there is precedent for the notion that subjects may have some strategic control over aspects of interactive activation processes. Rumelhart and McClelland (1982) showed that they could account for the effects of subjects' expectations for type of context in visual letter recognition studies (Carr, Davidson, & Hawkins, 1978) by assuming that subjects controlled the letter-to-word inhibition parameter in the interactive activation model of letter perception. Other types of strategic control over activation processes are under exploration in connectionist models of attention (Cohen, Dunbar, & McClelland, 1990), and differences in strategic control over such processes may be relevant to certain aspects of schizophrenic thought disorder (Cohen & Servan-Schreiber, 1992).

In any case, it appears that a stochastic interactive activation model which uses a straightforward representation of the experimental materials and which allows a reasonable amount of encoding failure can exhibit behavior just like that of human subjects. The model made blend errors with the appropriate frequency as the default assignment abilities of an interactive activation network would tend to fill in words from one overlap sentence when the trace for its pair was missing desired information (due to encoding failure). Surprisingly, the model also solved the mystery of the inconsistent conditional probability data, producing the observed pattern of recall performance not only in Experiment 1, whose data seemed to support synthesis-at-retrieval models, but also in Experiments 2 and 4, whose data pointed toward a single-trace account of postaccess interference rates. Because of the frequency of encoding failures suggested by the model, one can see that the probability of recalling a second word from a sentence, given the correct recall of a first word, will depend quite considerably on the probability that the second word was properly encoded in the first place. The particular probabilities found in the different conditions in each experiment can therefore be seen as a reflection of the particular composition of the sentence materials in that experiment, along with the encoding failure rate—all incorporated into the successful model.

Other Models

Given the success of our specific synthesis-at-retrieval model, it is worth considering how the two other classes of models would behave given similar assumptions about encoding failure. Let us first consider the single-trace models. There was no experimental evidence that subjects as a group consciously integrated similar sentences during the rating task, although the possibility remains that such trace synthesis occurred unconsciously or in a sufficiently fleeting manner, despite our manipulations to minimize it. If one wishes to believe in this latter possibility, then the ability of single-trace models to synthesize traces during encoding, shared with synthesis-at-retrieval and synthesis-in-storage models, would allow them to account for the apparent memory blends as more of an elaboration-at-encoding phenomenon.

However, it might also be the case that a single-trace model like Shiffrin's SAM model would also fit the data if multiple access attempts were allowed in cases where an initial trace selection resulted in incomplete recall, as might occur after encoding failure. Successive attempts could proceed just like the first, with the probability of selecting a given trace based on that trace's similarity to the cue. Because overlap sentence traces equally match the cue, they ould have an equally good chance for se-ction. So if a first attempt accessed an verlap sentence trace containing only one f the two target words, and a second at-empt accessed the other overlap sentence, e subject might make a crossover error. It eems plausible that this type of single-ace/multiple-access model could give an ccount for at least the general trends in the ata.

Nevertheless, this kind of extension of e single-trace model really amounts to rning it into a synthesis-at-retrieval ac-ount. This modified single-trace model nd our model essentially only differ in that e multiple traces are accessed in succes-ion in one case and simultaneously in the ther. Unfortunately, as with many theo-etical disputes between sequential and par-llel models, it would be difficult to devise n experiment to adequately test between e revised single-trace model and a synthe-is-at-retrieval model. One might measure esponse times to crossover errors as com-ared to correct responses, with the hy-othesis that successive accessing of mul-iple traces would lead to longer response mes in the case of the blends. But synthe-is-at-retrieval models such as the one pre-ented here would also predict longer re-ponse times with crossover errors, as they ndeed take longer to settle into a stable attern of activations when the extra com-etition between overlap sentence traces nd the lower "goodness" level of blends re involved.

As for the synthesis-in-storage models, it emains to be studied how models using su-erpositional matrix- or convolution-based torage (for example, those of Knapp & nderson, 1984; McClelland & Rumelhart, 985; or Metcalfe, 1990) would manage the ask of simulating the specific patterns of ecall performance found in our experi-nents. It will be of interest to discover whether these types of models could pro-luce *as few* blend errors as were obtained n our studies, given their tendency to blur he distinctions between traces. Perhaps through the use of larger vectors represent-ing sentence properties or highly differen-tiable "context" vectors, these models could manage to preserve an appropriate distance between the representations of overlap sentences. After all, the represen-tation of overlap sentence traces in our syn-thesis-at-retrieval model do reflect a certain type of synthesis-in-storage, in so far as they share the same property unit in every occurrence of overlap. Thus it seems that the success of a synthesis-in-storage model may hinge on the degree to which it is iso-morphic to a synthesis-at-retrieval model like the stochastic interactive activation model outlined here. The synthesis-in-storage and synthesis-at-retrieval classes of memory models share more characteristics with each other than with the single-trace models and could be expected to perform more similarly.

Conclusion

The subjects in the four experiments def-initely exhibited less memory blending than might have been expected following the predictions of some synthesis-at-retrieval models. However, the data from the exper-iments do appear to be consistent with syn-thesis-at-retrieval models when they in-clude: (1) a source of inherent variability; and (2) the possibility of encoding failure. The version of the McClelland (1981) syn-thesis-at-retrieval model investigated here provided a very good qualitative and quan-titative account for the heterogeneous—and initially puzzling—experimental find-ings.

Single-trace models could potentially ac-count for the experimental data by either assuming that unconscious or fleeting pro-cesses lead to occasional synthesis of traces during encoding, or that successive retrieval of single traces (assuming encod-ing failure) leads to synthesis at retrieval.

The data poses some challenges for cer-tain synthesis-in-storage models. It is not self-evident whether or not the superposi-tional storage of traces in these models

would allow for the relatively low fre-quency of blend errors obtained in the ex-periments. The effort to determine whether such models can be made to fit the data may lead to new constraints on synthesis-in-storage models, as has been the case for models of the synthesis-at-retrieval type. Conversely, it will be interesting to exam-ine whether the prototyping and generaliza-tion properties of the synthesis-in-storage models, a natural advantage of their ten-dency to blur distinctions between traces, can be captured with the synthesis-at-retrieval model that we have presented here. These complementary studies may take us closer to understanding what gov-erns the human ability to generalize well and yet preserve relatively distinct access to particular prior events.

APPENDIX A: POST-TEST QUESTIONNAIRE

1. In Part 1, you read 40 sentences and evaluated 5 content words per sentence. Did you at any point no-tice that some sentences used some of the same con-tent words as other sentences had used?
2. After about how many of the 40 sentences did you first notice this duplication of words?
3. About how many sentences would you say shared content words with other sentences?
4. If a sentence shared words, were these words always shared with just one other sentence, or did some sentences share words with more than one other sentence?
5. If a given content word was shared between more than one sentence, did it appear in only two sentences, or did some content words appear in more than two sentences?
6. When sentences shared some of the five content words, did they always share the same number of words?
7. When you would realize that a sentence shared words with a sentence that you read earlier, did you recall the earlier sentence?
8. Did you ever try to relate the subject matter of sentences that shared words, or try to combine the similar sentences into a single, more complex sen-tence?

APPENDIX B: DETAILS OF THE SIMULATION MODEL

Processing in the model begins by setting external inputs to the network and resetting the activation of all units to 0. Processing then proceeds for 100 cycles. In each cycle, net inputs to each unit are calculated based on existing activations, then activations are updated based on the net inputs. The net input to a particular unit i at time t is:

$$net_i(t) = \Sigma_j w_{ij} o_j(t) + ext_i(t) + \mathcal{E}_\sigma$$

where the summation ranges over all units j with connections to unit i,
w_{ij} is the weight between another unit j and unit i,
o_j is the output of unit j (which equals the activation a_j for all $a_j > 0$, and 0 otherwise),
$ext_i(t)$ is the external input to the unit (set to 1.0 for a cue), and
$\mathcal{E}_\sigma$ is normally distributed noise with a mean of 0 and standard deviation σ.

Once the net input to a unit is computed, the result-ing change in its activation depends on whether the net input is >0:

if $net_i > 0$,
$$\Delta a_i = istr(max - a_i)net_i - decay(a_i - rest),$$
otherwise,
$$\Delta a_i = istr(a_i - min)net_i - decay(a_i - rest),$$
where
istr is a parameter scaling the relative size of the influence of inputs to units,
max is the maximum activation parameter,
decay is a parameter determining the strength of the tendency to return to resting level,
rest is the resting activation parameter, and
min is the minimum activation parameter.

Finally, the resulting activation is time averaged be-fore choosing the most active unit:

$$\bar{a}_i(t) = \lambda a_i(t-1) + (1-\lambda)\bar{a}_i(t-1),$$
where λ is a parameter governing the amount of time averaging.

The values of all parameters were set to the default values used in McClelland and Rumelhart (1988) and McClelland (1991):

$\sigma = 0.025$;
$istr = 0.1$;
$decay = 0.1$;
$max = 1.0$;
$rest = -0.1$;
$min = -0.2$;
$\lambda = 0.05$.

REFERENCES

ANDERSON, J. R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Lawrence Erlbaum Asso-ciates.

ANDERSON, J. R. (1983). *The Architecture of Cogni-tion*. Cambridge, MA: Harvard University Press.

ANDERSON, J. R., & BOWER, G. H. (1971). On an as-sociative trace for sentence memory. *Journal of*

Verbal Learning and Verbal Behavior, 10, 673–680.

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

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

Carr, T. H., Davidson, B. J., & Hawkins, H. L. (1978). Perceptual flexibility in word recognition: Strategies affect orthographic computation but not lexical access. *Journal of Experimental Psychology*, 4, 674–690.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97, 332–361.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99, 45–77.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1–65.

Grossberg, S. (1976). Adaptive pattern classification and universal recording: Part 1. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23, 121–134.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1). Cambridge, MA: MIT Press.

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

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96, 208–233.

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 616–637.

Loftus, E. F. (1977). Shifting human color memory. *Memory and Cognition*, 6, 696–699.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 19–31.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Meeting of the Cognitive Science Society* (pp. 170–172).

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23, 1–44.

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

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114, 159–188.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing*. Cambridge, MA: MIT Press/Bradford Books.

McCloskey, M., & Zaragoza, M. (1985a). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, 114, 1–16.

McCloskey, M., & Zaragoza, M. (1985b). Postevent information and memory: Reply to Loftus, Schooler and Wagenaar. *Journal of Experimental Psychology: General*, 114, 381–387.

Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review*, 89, 627–661.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General*, 119, 145–160.

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

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

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89, 60–94.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel Distributed Processing* (Vol. 1, pp. 25–31). Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 2.), Cambridge, MA: MIT Press.

(Received May 7, 1991)
(Revision received December 30, 1991)