Journal of Experimental Psychology:
Human Learning and Memory

VOL. 6, NO. 4 JULY 1980

Readability and Recall of Short Prose Passages: A Theoretical Analysis

James R. Miller and Walter Kintsch
University of Colorado

Readability can be viewed as an interaction between a text and the reader's prose-processing capabilities, rather than as some innate property of a text. In support of this view, this article applies an extended and formalized version of the Kintsch and van Dijk prose processing model to 20 texts of varying readability. Each text was read by 120 subjects; reading times and recall protocols were collected. In addition, each text was analyzed by the processing model. This analysis allowed recall predictions based on the frequency of microstructure processing of a text's propositions and readability predictions from the frequencies of events that indicate processing difficulty. Recall predictions were moderately successful in spite of the absence of macrostructure processing, and multiple regression predictions of readability (reading time per proposition recalled), reading time, and recall ranged from $r = .8$ to .9.

Readability has been and continues to be a problem of considerable practical interest. Historically, researchers have approached this problem with the assumption that a text's readability could be measured as a simple function of text features that are objective and easy to determine, but that are also quite superficial (e.g., sentence length and word frequency). This approach has been only moderately successful in practice, and it provides very little of interest for the theorist concerned with reading comprehension (see the review by Kintsch & Vipond, 1979).

An alternative approach is developed here that regards readability as the result of an interaction between the text and the reader. A model will be proposed that simulates certain aspects of the comprehension process, in particular, the maintenance of the text's semantic coherence (Kintsch & van Dijk, 1978). We hypothesize that at those points in the comprehension process at which the model has difficulty locating and maintaining coherent relations, human readers should experience similar difficulties. Hence, the performance of the

This research was performed within the Institute for the Study of Intellectual Behavior, University of Colorado, and is Publication No. 88 of the institute.

The research was supported by Grant NIE-G-78-0172 from the National Institute of Education and Grant MH-15872 from the National Institute of Mental Health.

Computer time for the model construction was provided by the SUMEX-AIM computer facility under Grant RR-00785 from the National Institutes of Health and by the Computer Laboratory for Instruction in Psychological Research, which is supported in part by the University of Colorado.

We thank Ely Kozminsky, who collected the data, and Donna Caccamise who supervised the scoring of the recall protocols.

Requests for reprints should be sent to James R. Miller, Department of Psychology, University of Colorado, Boulder, Colorado 80309.

 0096-1515/80/0604-0335$00.75

model on a given text should predict how difficult a homogenous group of readers should find a text.

Comprehension difficulties may express themselves in two ways. First, a difficult text may require additional processing to maintain coherence, and so should require extra time for reading. Second, if these necessary additional processes are not performed, the representation of the text will be deficient, and recall should suffer. We suggest, therefore, that the best index of readability is a measure that takes both factors into account, that is, reading time per unit recalled. Since readability predictions involve both reading times and amounts of recall, they provide sensitive measures for the evaluation of a model of text comprehension. Our aim in this report is to outline such a model, and to test it via its recall and readability predictions for a number of texts.

Comprehension encompasses more than understanding words or phrases or decoding propositions. These elements of understanding must also be arranged in a coherent pattern. For instance, scrambling the sentences in an otherwise readable text significantly impairs comprehension (e.g., Thorndyke, 1977). Other possible sources of coherence disruption have been suggested by Kintsch and Vipond (1979), within the framework of Kintsch and van Dijk's (1978) model of prose comprehension. This model asserts that as the reader works through a text, only a fraction of the already read text can be held in short-term (working) memory. If a segment of text is read that is not related to the current contents of working memory, long-term memory must be searched to locate a part of the text that can interrelate what has been read previously with the current input segment. If this search is successful, that part of the text is reinstated in working memory to maintain the coherence of the text.

This search will not always be successful: A segment of text may be encountered that bears no explicit connection with what has already been read. This could be due to a major topic shift in the text, the author's carelessness, or the improper segmentation of the text by the reader. When coherence fails in this way, the reader must generate a connecting or bridging inference that will connect this segment with the preceding text. The occurrence of these knowledge-based inferences has been amply demonstrated (e.g., Haviland & Clark, 1974; McKoon & Keenan, 1974), and they should be a source of reading problems to the extent that the reader lacks relevant knowledge or inferencing strategies.

Kintsch and Vipond (1979) have offered some ad hoc analyses suggesting that these sources of coherence disruption are in fact related to reading difficulty. We shall attempt a test of these ideas here. Specifically, this test will be concerned with only one part of the Kintsch and van Dijk (1978) model, that is, the formation of the microstructure of a text's propositions and its implications for readability. This is a simplification, since macroprocesses undoubtedly affect readability. However, Kintsch and van Dijk (p. 388) found some indications that when people read short paragraphs out of context for purposes of immediate recall, their recall primarily reflects microprocesses. Hence, we shall work with short prose paragraphs and an immediate recall task in an effort to isolate and test the microprocessing component of the model. The success of the model's recall predictions will thus indicate the importance of macroprocesses in this isolated task: If macroprocesses are relevant in this task, the model's recall predictions should be impaired.

The most salient finding in prose memory is what has come to be called the *levels effect*: People remember the "important" information in a text much better than the "lower level" details. Although this effect is experimentally well established (e.g., Caccamise & Kintsch, 1978; Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; Mandler & Johnson, 1977; McKoon, 1977; Meyer, 1975; Meyer & McConkie, 1973; Thorndyke, 1977; Waters, 1978), its theoretical significance is less clear. At the most basic level, there are atheoretical demonstrations that those portions of a text rated most "important" by subjects are also best recalled (Johnson, 1970) and recognized (Caccamise & Kintsch, 1978). A

number of more sophisticated structural analyses of text have also been shown to correlate with recall, such as Kintsch's (1974) propositional networks, Meyer's (1975) hierarchical structures, and the story grammars of Mandler and Johnson (1977) and Thorndyke (1977). These analyses are not always directly comparable, however, since some analyses focus on the macrostructure of texts (e.g., story grammars), whereas others are concerned with much smaller text units (e.g., propositions).

Another problem with structural models is that they have neglected the relationship between the levels effect and the processes operating on these structures. Hence, this research also considers the process-based explanation for the levels effect offered by Kintsch and van Dijk (1978); that is, the important propositions are remembered better because they are processed more frequently. This explanation would have the advantage of being based on the interaction between the structure and processes of the prose comprehension systems. Since this explanation—as well as the readability predictions—is dependent on the exact nature of the processing model, we should describe those aspects of the Kintsch and van Dijk model that are relevant to the present study, the elaborations of that model made in this research, and the formalization of the model as a computer program.

Kintsch and van Dijk (1978) adopted a modular approach to text comprehension. Comprehension is decomposed into several major component processes: (a) the initial parsing of the written or spoken text into a conceptual (propositional) representation, (b) the arrangement of these propositions into a coherent structure called the text base, (c) the use of world knowledge to organize individual elements of the text base into global concepts, and (d) the construction of the text's macrostructure. The organization of the text base and the development of a macrostructure were the components systematically developed in the original model, whereas natural language parsing and the use of world knowledge were neglected. The present study is primarily concerned with the construction of a coherent text base.

Since the basic model has been described in detail in Kintsch and van Dijk (1978) and Kintsch and Vipond (1979), we shall merely outline it here, noting our recent changes and elaborations. Briefly, the short-term memory allocation strategy (the *leading edge* strategy) has been modified somewhat, and an explicit system has been added to select appropriate groups of propositions during input. The model described here exists as a computer program written in Interlisp in two major parts, a chunking program that performs the initial segmentation of the text and a microstructure coherence program that simulates the comprehension process and yields the information on which the recall and readability predictions are based. In addition, a chi-square minimization program is used to fit the actual data to the model's predictions and provide parameter estimates.

As noted earlier, the model does not include a parser to derive the propositional representation from the verbatim text. This significant step is, at present, bypassed; the text is hand coded into an ordered list of propositions. This procedure is not arbitrary and is based on some explicitly stated linguistic principles, but it is by no means completely objective (for details, see Kintsch, 1974; Turner & Greene, 1978). The following is an example of a short text (one of the paragraphs ["Saint"] used in the experiment); its propositional representation is shown in Table 1.

In the request to canonize the "Frontier Priest," John Newmann, bishop of Philadelphia in the 19th century, two miracles were attributed to him in this century. In 1923 Eva Benassi, dying from peritonitis, dramatically recovered after her nurse prayed to the bishop. In 1949 Kent Lenahan, hospitalized with two skull fractures, smashed bones, and a pierced lung after a traffic accident, rose from his deathbed and resumed a normal life after his mother prayed ardently to John Newmann.

The propositions are ordered by the appearance in the text of the words corresponding to their predicates and are labeled for convenience. Propositions embedded as arguments in other propositions are referred to by their label (e.g., P1).

The chunking program. A critical assumption of the model is that readers

Table 1
The Propositional Text Base for "Saint" Sample Paragraph Prior to Segmentation by the Chunking Program

Number	Proposition	Number	Proposition
(P1	(REQUEST P2 P8))	*SENTENCE*	
(P2	(CANONIZE P3))	(P16	(TIME:IN P17 1949))
(P3	(ISA JOHN-NEWMANN FRONTIER-PRIEST))	(P17	(HOSPITALIZED KENT-LENAHAN P18 P20 P21))
(P4	(ISA JOHN-NEWMANN BISHOP))	(P18	(FRACTURES SKULL KENT-LENAHAN))
(P5	(LOC:IN P4 PHILADELPHIA))	(P19	(TWO P18))
(P6	(TIME:IN P4 19TH-CENTURY))	(P20	(SMASHED BONES KENT-LENAHAN))
(P7	(TWO MIRACLES))	(P21	(PIERCED LUNG KENT-LENAHAN))
(P8	(ATTRIBUTED P7 JOHN-NEWMANN))	(P22	(AFTER P17 ACCIDENT))
(P9	(TIME:IN P8 THIS-CENTURY))	(P23	(TRAFFIC ACCIDENT))
SENTENCE		(P24	(ROSE KENT-LENAHAN DEATHBED))
(P10	(TIME:IN P11 1923))	(P25	(RESUMED KENT-LENAHAN P26))
(P11	(DYING EVA-BENASSI PERITONITIS))	(P26	(NORMAL LIFE))
(P12	(DRAMATICALLY P13))	(P27	(AFTER P25 P28))
(P13	(RECOVERED EVA-BENASSI)	(P28	(PRAYED MOTHER JOHN-NEWMANN))
(P14	(AFTER P15 P13))	(P29	(ARDENTLY P28))
(P15	(PRAYED NURSE BISHOP))	*SENTENCE*	

Note. See appendix.

process a text in a number of cycles. A limited number of propositions are entered into the model and interrelated with propositions encountered during previous cycles. Kintsch and van Dijk (1978) stated that chunks are formed at syntactic boundaries, that is, at sentence and major phrase boundaries. They did not, however, rigorously define how these boundaries were to be identified. In line with the formal semantic emphasis of the model, we have defined a procedure that determines appropriate text segments by applying a set of simple heuristics to the semantic information in the proposition list and to the text itself. This procedure describes a consistent method for identifying reasonable subsentence text segments, based on properties of both the text and a hypothetical reader.

The chunking program operates by reading, in order, a word from the text (disregarding function words) and locating the corresponding concept in the proposition list. It then compares this proposition to those already in the current chunk to determine whether the proposition should be added to this chunk, or if the existing chunk should be terminated and a new one begun with this proposition. In addition to the primary rule that a sentence boundary always terminates a chunk, six chunking heuristics are used.

1. A chunk must have at least two propositions.

2. If all of the arguments of the located proposition are themselves propositions (as in Proposition P1 in Table 1), include it in the current chunk only if all the propositions so referenced are already part of this chunk.

3. If the located proposition's arguments have already been read, include the proposition in the chunk.

4. Propositions can be embedded, or used as arguments, in other propositions. For example, in the sample paragraph, Proposition P2 is embedded in Proposition P1. If the located proposition is embedded in one of the propositions already included in this chunk (or vice versa), include this proposition in the chunk.

5. If the located proposition shares arguments with the proposition immediately preceding it, include it.

6. If all of the preceding rules fail, terminate the current chunk and begin a new chunk with the located proposition.

There are two additional considerations. First, these rules are applied in the order presented, that is, Rule 2 is not considered until Rule 1 is tested and found not to apply. Second, the reader being simulated by this program is assumed to have a minimum number of words that will always be read before a chunk is formed. This value, *I*,

includes both function and content words, and is a free parameter of the model.

The sample paragraph shown earlier is broken down into four chunks with $I = 19$. This value yields the best recall and readability predictions for that paragraph; the selection of this value will be described later. The first two chunks are defined by the first two sentence boundaries. The third chunk is terminated after Proposition P26, just before the word "after" in the last sentence of the text; this segmentation is due to the execution of Rule 2. The fourth chunk contains the remaining propositions of the third sentence. Notice that if I had been smaller, more chunks would have been formed. For instance, with $I = 9$, seven chunks are formed, beginning with propositions P1, P7, P10, P14, P16, P24, and P27. Note also that these are all intuitively reasonable boundaries for sentence segmentation.

The important points about the chunking program are its procedure for formally defining phrase boundaries and its use of the actual text in conjunction with the propositional representations. One drawback to the system is its limited syntactic knowledge, using only sentence boundaries as aids to chunking. We do not doubt that humans can and do employ some syntactic guidelines when segmenting a text, but, for our present purposes, the simple heuristics described here seem sufficient. Problems could be expected with paragraphs consisting of many very short sentences: The current chunking program would segment each sentence into a separate chunk. This problem might be avoided by making the sentence boundary check secondary to the requirement for the reading of a minimum number of words before a chunk can be created.

The microstructure coherence program. This program is essentially a formalization of the microstructure portion of the Kintsch and van Dijk (1978) model, although some minor modifications have been made. The program uses as input a proposition list already segmented by the chunking program.[1] The goal of the coherence graph system is to add a chunk of propositions to the current contents of the short-term buffer. The buffer is empty at the beginning of the first cycle, and a superordinate proposition must therefore be selected to stand at the head of the graph. We should stress the importance of this selection process —placing a nonoptimal proposition at the head of the coherence graph can lead to highly inappropriate processing and incorrect predictions. At present, this selection has to be done on an intuitive basis because of the lack of well-defined macroprocesses in the model. Eventually, we hope to develop a model in which the choice of the superordinate microproposition is determined by the model's macroprocesses.

Given the superordinate proposition, the program adds the chunks remaining to the text base by building a coherence graph in working memory. This is done by connecting the other propositions to the superordinate on the basis of argument overlap, the coherence criterion used by the model. All propositions that share arguments with the superordinate or are embedded in the superordinate are placed at Level 2 of the graph and are connected to the superordinate. The propositions from this input chunk as yet unassigned to the graph are then compared to the propositions now at Level 2, and any that share arguments or are referenced by the Level 2 propositions are so connected and placed at Level 3. If there is more than one proposition at Level 2 to be compared to the input propositions, the comparison takes place in order of the recency of the Level 2 propositions. This matching procedure is repeated until (a) all propositions from this chunk are entered into the graph or (b) some propositions remain that cannot be added to the graph because of the lack of common arguments. In addition, a long-term memory graph is constructed by a similar process, beginning

[1] The sequential operation of the chunking and coherence programs was done primarily for our convenience and to increase the efficiency of the coherence program. This was possible because the two programs are theoretically and computationally independent. It would be possible to merge the two programs so that a segment of text is located by the chunking algorithm and passed on-line to the input stage of the coherence program, but this is only a computational detail with no theoretical impact.

with the same superordinate proposition. This long-term graph will ultimately have added to it all of the propositions of the text, according to the principle of argument overlap.

If all the propositions of the input cycle are successfully added to the working memory graph (as is the case in Cycle 1 in the appendix), a subset of these propositions is selected to remain in the buffer for processing on the next cycle. This selection will be described later. If some propositions cannot be added to the buffer because of the lack of argument overlap, the model must take steps to correct the coherence failure. Two possible conditions exist. The remaining propositions could be related to something that had been read previously but that is no longer present in working memory. Alternatively, these propositions could be part of a new topic for which no previous reference exists in the text.

In these cases, the long-term memory graph is searched for a proposition that can interconnect the propositions already in the buffer with those as yet unconnected. This search begins at Level 1 of the long-term graph. If no such linking proposition can be found, it is assumed that a bridging inference is made to maintain the coherence of the text. The model does not specify what that inference is, but merely notes that an inference is required. A new graph is then constructed in working memory from the propositions not yet connected to the initial graph. If, however, a linking proposition can be found, that proposition is reinstated in working memory, and a new graph is constructed from the linking proposition and the propositions that could not be connected to the previous buffer. Note that in both of these cases, the construction of a new graph in working memory requires that one of the propositions be specified as the superordinate of the graph.

When the coherence graph construction is completed, some of the propositions in the graph are selected to remain in the buffer during the next cycle. Thus, some propositions are processed in more than one cycle. This selection is accomplished by a slightly modified version of the leading edge strategy described by Kintsch and van Dijk (1978) and Kintsch and Vipond (1979). The short-term memory buffer has a basic capacity of s propositions, which are selected from those currently available as follows: (a) retain the superordinate proposition; (b) retain the most recent Level 2 proposition that is pointed to by the superordinate; (c) continue this process from Level 2 to the bottom of the graph; and (d) if more propositions may yet be added to the buffer, add propositions from Level 2 in order of recency, and likewise with the remaining levels of the graph. All of these processes halt, however, when the limiting number of s propositions has been added to the buffer.

This is the leading edge strategy as it was used before. In working with the model, however, we found that the model's recall predictions could be improved by increasing the flexibility of the buffer. First, the size of the buffer is expanded from s to $s + 1$ slots during Cycle 1, reflecting the less complete use of the reader's processing capacity early in the text. The view of short-term memory as a flexible buffer, with a capacity that depends at least to some extent on the resources that are available for maintaining it, is derived from current memory research. For instance, Kintsch and Polson (1979) have argued that in list-learning tasks, the buffer is quite large while the subject is processing the first few items of the list, but it shrinks considerably as the learner's processing resources become overloaded. By permitting the buffer to be somewhat dependent on resource availability, we have tried to account for this factor.

Second, the propositions selected for retention in the buffer by the leading edge rule are determined not only by their level and recency in the buffer but also by their content. Specifically, when a proposition is selected for retention, it is inspected to determine whether any of its arguments are embedded propositions. If so, those propositions are immediately selected for retention. This embedding rule avoids some intuitively unlikely buffer configuration (e.g., that a reader would retain from one cycle to another that "John did something" but would not retain the embedded proposition describing what John did). Thus, the buffer may contain a maximum of $s + 1$

propositions if all of those propositions are associated by embedding relations. This "stretching" of the buffer is typically applied when working with a very complex and heavily embedded text. It seems reasonable to assume that these occasions should be most frequent in texts that are difficult to read, so that short-term memory overloads might be expected to be correlated with readability.

The appendix shows how the proposition list shown in Table 1 is processed by the coherence graph program. The sentence and chunk boundaries were generated with $I = 19$; the buffer size parameter s was set to 2.

In Cycle 1, P1 is specified as the superordinate proposition, and Propositions P2, P8, and P9 are connected to it at Level 2 (P2 and P8 are embedded in P1, and P1 and P9 both reference P8). Level 3 consists of P3, P4, and P7, connected to P8, and Level 4 contains P5 and P6, connected to P4.

Since this is Cycle 1, the leading edge rule is applied to the graph with an expanded s of 3. The superordinate Proposition P1 is selected for retention; embedded within it are P2 and P8. In addition, P2 and P8 have P3 and P7 embedded within them, respectively. Thus, simply on the basis of the embedding rule, five propositions should be retained, whereas on this cycle, s is only 3. Since all of these propositions are related by embedding relations, the buffer can be expanded to one more than its specified size, and four of the five propositions can be retained. P3—the lowest level, least recent proposition—is deleted,[2] and the short-term buffer will contain a graph of P1, P2, P8, and P7 at the beginning of Cycle 2.

In Cycle 2, coherence breaks down; None of the propositions can be connected to those in the buffer. This initiates a long-term memory search that fails for P10, P11, P12, P13, and P14 but succeeds with P15, which matches P4 through the argument BISHOP. Hence, P4 is reinstated and a new graph is constructed with P13 selected to be at the head of the graph. A long-term memory search is also necessary in Cycle 3, except that no linking propositions can be found this time, indicating that a bridging inference must be generated. The occurrence of this inference is noted, and a new graph is built with P18 at the head. The processing of Cycle 4 proceeds smoothly, although the operation of the leading edge rule might be noted. Although P18 is at the head of the graph, none of the propositions in this input chunk contacted it; graph construction was centered on P25. This was a frequent occurrence in the analysis of the texts in these experiments, as the local topic shifted throughout the course of the text. Retaining P18 at this point would impede the representation of this topic shift, so the leading edge rule was formulated to omit superordinate propositions if they were not contacted by any proposition from the current input chunk. This rule emphasizes the buffer's role as the representation of the *local* focus of the text; the global representation is more properly a concern of the text's macrostructure.

Thus, for any given text (or rather, the corresponding proposition list as segmented by the chunking program), one can specify values of s and I and simulate the processing of that text by means of the present program. A number of relevant processing statistics can then be extracted from the simulation protocols (see appendix). First, the number of times each proposition was processed is recorded; these frequencies are fitted to the recall data with the chi-square minimization program. Other processing statistics, in particular the number of inferences and reinstatements made while processing the text, are used to predict the readability of the text.

Method

Subjects

Six hundred subjects enrolled in an introductory psychology course at the University of Colorado participated in the experiment in fulfillment of a course requirement. Each subject read and recalled a practice paragraph and four experimental texts, as well as some material not related to the present study.

Materials

Twenty paragraphs served as the experimental material. These paragraphs were selected from

[2] It might alternatively be argued that propositional embedding manifests short-term memory chunking, and that therefore all embedded propositions should be retained (Bjork, Note 1). There are too few such cases in our data to permit a choice between these alternatives.

Table 2
Descriptive Statistics for the 20 Test Paragraphs

Statistic	Low	High	*M*	*SD*
Text length (words)	67.0	85.0	77.0	4.14
Sentence length	12.0	42.0	20.0	7.52
Propositions/paragraph	24.0	33.0	29.0	2.85
Proposition density[a]	2.0	3.4	2.7	.34
Arguments/paragraph	11.0	21.0	17.5	2.65
Mean word frequency[b]	60.0	359.0	208.0	77.23
Flesch score[c]	8.7	80.6	50.2	18.26
Subjective readability[d]	2.2	3.6	2.7	.49
Subjective interest[d]	1.4	4.5	3.3	.76

[a] Words per proposition.
[b] Based on Kučera and Francis (1967) norms.
[c] Flesch (1948).
[d] Subjective ratings were collected from an independent group of 16 subjects. A rating of 1 represented a text of high subjective readability or interest.

Reader's Digest, with the restriction that they could be reasonably well understood when read outside of their original context. Table 2 presents a number of descriptive statistics for the selected texts.

Procedure

The 600 subjects were randomly assigned to five groups of 120 subjects. Each group was assigned four experimental texts to read. The order in which these texts were read was counterbalanced over subjects.

Each subject was tested in an experimental session that lasted up to 1 hr. Materials and instructions were presented on a computer controlled screen at the CLIPR facility of the University of Colorado.

Subjects read one practice paragraph that appeared on the screen. The same practice paragraph was used for all subjects. The subjects were instructed to read the text at their own rate, and to press a response button when they were finished; this button recorded their reading time. Subjects were then asked to recall in writing as much of the paragraph as they could, although not necessarily verbatim. There were no time restrictions, and subjects could change and add to their protocols as much as they liked. The four experimental paragraphs were then read and recalled in the same way.

Results[3]

Protocol Scoring

Copies of the 2,400 protocols were collected and scored against the propositionalized form of the corresponding paragraph. A lenient scoring criterion was adopted so that a proposition in the text base was scored as recalled if any meaning-preserving paraphrase was present. As is typical of immediate recall of short paragraphs, almost all of the protocols could be scored as reproductive recall; reconstructions, inferences, metastatements, and unrelated statements formed a very small portion of the protocols and were not further analyzed. The protocols were scored independently by two judges, one of them with previous experience in propositional scoring of recall protocols. Each judge scored 10 of the paragraphs. In addition, a subset of 230 protocols was scored by both judges, with a resulting reliability of .92.

Text Differences

Attempting to predict differential levels of recall, reading time, and readability among the 20 experimental texts would be pointless unless statistically significant differences in the texts can be demonstrated. The differences of the texts on these variables were highly reliable. In these texts, mean reading time varied from 51.1 to 85.0 sec, with an overall average of 68.3 sec, $F(15, 1785) = 59.97$, $MS_e = 497.06$. Mean number of propositions recalled ranged from 15.1 to 22.6 and averaged 18.8 propositions, $F(15, 1785) = 22.95$, $MS_e = 13.88$. The average value of the readability measure was 3.7 sec per proposition recalled, with a range of 2.5 to 5.1 sec, $F(15, 1785) = 19.55$, $MS_e = 4.12$.

There was a small but statistically significant effect of text order: Recall increased from Serial Position 1 to 4 by an average of 1.5 propositions, $F(3, 595) = 13.82$, $MS_e = 20.30$. However, the interaction between serial position and the five text groups was not significant, and more importantly, essentially the same pattern of recall was observed in Serial Positions 1 and 4 across the propositions in a given text. The frequency of recall of a proposition when its text appeared in Serial Positions 1 and 4 correlated .85. Thus, even though an overall serial position effect was present, the pattern of recall appeared to be independent of serial

[3] All analyses of variance were tested with $\alpha = .01$.

position. Hence, there was no reason why the data from different serial positions could not be pooled for further analyses.

Model Analyses: Recall

The propositional text bases for the 20 experimental texts were first analyzed by the chunking program to determine the text segmentations for various values of the input size parameter I. A lower boundary of I was set at 6 words; for each text, all larger values of I were explored from that boundary up to the number of words in the longest sentence of the text. At this upper bound, the text was segmented only by sentence boundaries. Because some paragraphs consisted entirely of very short sentences, whereas others contained some very long sentences, the number of possible input sequences varied greatly among paragraphs. For one text, the minimum I value of 6 resulted in segments terminated solely by sentence boundaries, whereas for another paragraph, as many as 10 different segmentations were possible, with I ranging from 6 to 30.

No formal evaluation was performed on the adequacy of the chunking program's performance. However, the model always formed chunks that were intuitively reasonable. Although it did not base its decisions on any syntactic knowledge, major phrase boundaries were generally respected.

On the basis of previous experience with the model, we decided to investigate values of s ranging from 1 to 5.[4] Thus, for each paragraph, the data were used as a standard to test the predictions of the model configurations obtained from the combination of five s values and a variable number of I values. A total of 560 simulations were needed to test all parameter combinations for all 20 paragraphs. Each simulation produced a detailed protocol showing how the model processed that text base for the given parameter combination; the appendix gives an example of such a protocol.

The recall predictions are based on the assumption that each time a proposition is processed by the model, the likelihood that it will be reproduced on a subsequent recall test increases by some predictable amount. All propositions are processed once on input, and so should be recalled at least some portion p of the time. However, some propositions are selected by the leading edge strategy to be processed in more than one cycle, and others are reinstated as the result of a long-term memory search. In general, if a proposition is processed n times, it should be recalled with a probability of $1 - (1 - p)^n$.

Hence, the model predicts a constant level of recall for these propositions that are processed only once—when they are first read—and increased recall for those propositions that are retained in the buffer for extra cycles and so receive additional processing. However, the recall protocols revealed that some propositions were recalled with probabilities far below the general level of recall of their paragraph. If we assume that all propositions are read by the subjects, and so should be recalled at least at this modal level, some provision must be made for those propositions that are seemingly omitted during recall.

When the roles of these low-recall propositions in their paragraphs were examined, we found that most of them could be generated by reconstruction rules as described by Kintsch and van Dijk (1978). For instance, if it is known, as in the earlier sample paragraph, that Eva recovered from a fatal illness because of a nurse's prayer, it can plausibly be assumed that the recovery was perceived as "dramatic" (P12). However, there are other such elaborations that could have been included in the text, but happened not to be—perhaps that Eva was expected to die *at any moment*. Since the subjects were to recall the paragraphs without these potentially valid extrapolations, the difficulty of discriminating old and inferred statements (Bransford & Franks, 1971; Keenan & Kintsch, 1974; Sachs, 1967) would lead the subjects to suppress *all* elaborations. To prevent such reconstruc-

[4] In the few cases in which minimum chi-square values were obtained with $s = 5$, larger s values were explored to determine the actual minimum. However, the larger values always resulted in poorer estimates.

tion suppression strategies from distorting our parameter estimates, any redundant proposition that was recalled with a frequency of 1 *SD* or more below the mean recall for all propositions in the text in which they appeared was excluded from the parameter estimation analyses. Only redundant propositions were excluded; over all the 20 texts in the experiment, 85 propositions (15% of the total) were excluded by this rule. Excluding them seemed preferable to expanding the model to include an ad hoc "redundancy" parameter reflecting the status of these propositions. Instead, this exclusion method clearly distinguishes those trends in the data that the model predicts from those that are merely explained in a general way.

The observed and predicted recall frequencies for the resulting proposition list were submitted to a parameter estimation program employing a minimum chi-square criterion (Wickens, Note 2). This program is a hill-climbing procedure, similar in function to Chandler's (Note 3) STEPIT. One parameter, the reproduction probability p, was estimated by this procedure. A total of 485 data points were available from the 20 paragraphs, which contained from 19 to 29 propositions following the exclusion of redundant propositions. Each data point was based on the recall of 120 subjects. The observed (and predicted) recall frequencies for the sample paragraph are shown in Table 3.

The sum of the minimum chi-squares for the 20 paragraphs was $\chi^2(465) = 1{,}411$. Although this value indicates that the deviations between the best possible predictions of the model and the data are highly significant statistically, the absolute values of the chi-squares for individual texts are not excessive, given that the use of 120 subjects results in an extremely powerful test. The average chi-square per paragraph was 70.5 for 23 degrees of freedom; 6 of the paragraphs yielded nonsignificant values, 9 were statistically significant but not excessive (ranging from 50 to 90), and 5 paragraphs were fit poorly, with chi-squares between 108 and 143.

The values of I and s corresponding to the minimum chi-square estimations were distributed over their possible range. Although the variation of I appeared unsystematic, there was a clear tendency for low chi-squares to be associated with s values of either 1 or 2. We therefore explored the possibility of restricting the parameter space to low values of s; this would be in agreement with the low s values found by Kintsch and van Dijk (1978) and Spilich, Vesonder, Chiesi, and Voss (1979). The ratio of the chi-square for the restricted parameter space and the absolute minimum chi-square can be computed, and will be approximately distributed as F values, permitting a test of the feasibility of restricting the values of s. This is a rather sensitive test, since $F(465, 465) = 1.18$ at $\alpha = .05$. This test rejects the

Table 3
Observed and Predicted Recall Frequencies for the "Saint" Sample Paragraph

Proposition	Data	Model	χ^2 [a]	Proposition	Data	Model	χ^2 [a]
P1	101	101.71	.00	P16	77	73.15	.20
P2	102	101.71	.00	P18	93	101.71	.75
P3	90	73.15	3.88	P19	63	73.15	1.41
P4	93	101.71	.75	P20	74	73.15	.01
P5	86	73.15	2.26	P21	86	73.15	2.26
P6	60	73.15	2.36	P22	80	73.15	.64
P7	110	101.71	.68	P23	56	73.15	4.02
P8	104	101.71	.05	P24	59	73.15	2.74
P10	82	73.15	1.07	P25	98	112.86	1.96
P11	69	73.15	.24	P26	97	101.71	.22
P13	114	101.71	1.49	P27	104	101.71	.05
P14	111	101.71	.85	P28	84	101.71	3.08
P15	83	101.71	3.44	—	—	—	—

Note. P9, P12, P17, and P29 were excluded (see text).
[a] $\chi^2(24) = 34.40$, $p = .078$.

hypothesis that s can be restricted to any one value between 1 and 5 without resulting in a significant deterioration in the goodness of fit. However, if s is restricted to values of 1 or 2, a nonsignificant F value of 1.09 is obtained. Therefore, this restriction was used in all further analyses.

There is little evidence that would allow restricting the range of I. For several paragraphs, a variety of I values fit the data almost equally well, with no clearly defined minimum. It is possible that input size is best viewed as a subject-dependent parameter, and that when the data from 120 subjects are pooled, no clear picture can emerge. For one paragraph, a minimum chi-square of 56 was found for $I = 10$, but chi-square values of 61 and 59 were found for I values of 6 and 23, respectively. The recall data for this paragraph (and at least three others) obviously do not sufficiently constrain the estimation of I.

However, which of these nearly equal I values is ultimately selected for a given passage is very important to the readability predictions, since the number of reinstatements carried out during the construction of the coherence graph depends crucially on I. We have therefore adopted a joint criterion for selecting the best-fitting parameter combination; this criterion maximizes the accuracy of both the recall and readability predictions. A slight increase in chi-square is acceptable if a notably better readability prediction can be obtained. Hence, a region of acceptable chi-squares was constructed by placing a .999 confidence interval around each text's observed minimum chi-square. From this region of acceptable I values, one value was selected such that the frequencies of reinstatements and inferences—our a priori prediction of the variables most likely to affect readability—were maximally correlated with the readability measure of reading time per proposition recalled. In 14 of the 20 paragraphs, this value coincided with the actual minimum chi-square. In the remaining 6 paragraphs, improved readability predictions were obtained at the cost of a modest increase in chi-square. This increase was not significant with respect to either the set of absolute minimum chi-squares, $F(465, 465) = 1.16$, or the set of minimum chi-squares for s values of 1 or 2, $F(465, 465) = 1.06$. The summary characteristics of the final estimates are shown in Table 4.

Table 4
Summary Statistics for the Selected Parameter Estimates

Statistic	M	Low	High
Buffer size (s)	1.50	1	2
Minimum number of words per chunk (I)	14.75	6	29
Probability of proposition reproduction (P)	.65	.49	.83
Number of reinstatements	1.00	0	3
Number of inferences	1.00	0	2

Further measures of the model's recall predictions are available by comparing the performance of the model with a mean recall model which asserts that each of a text's propositions will be recalled at the level of the mean recall of the text's propositions. This model might be viewed as expressing the null hypothesis that text recall requires no considerations of structure or process. Fifteen of the 20 texts were fit best by the coherence model, a difference that is significant by a sign test ($p < .005$). The F ratio formed by summing the chi-squares generated by each of the two models across all stories also supported the coherence model, $F(495, 495) = 1.23$, $p < .05$. We had expected the difference between the models to be greater. However, the data fitting was limited by the proximity of some of the recall data to ceiling performance. Mean recall across texts was 71%, with performance on 6 of the texts exceeding 75%. A ceiling effect would have two effects on our model fitting: It would clearly benefit the model predicting the mean because the range of propositional recall levels are increasingly flattened by the ceiling, and it would also result in a less sensitive test of the coherence model, as the differences between the various levels of propositional recall predicted by the model become reduced.

The mean recall model was presented for statistical comparisons with the coherence model; it cannot be seriously considered due to its prediction of zero correlation between observed and expected recall frequencies. The coherence model's

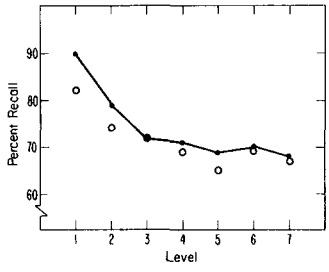

Figure 1. Recall as a function of the level of a proposition in the text base. (Open circles = data, closed circles = model predictions.)

correlation between observed and predicted recall frequencies over all paragraphs was significant ($r = .60$). For individual paragraphs, the observed–expected correlation was significant for 16 of the 20 paragraphs. Inspection of the remaining 4 texts showed that these generally have a listlike structure; for example, collections of the properties of the abominable snowman, of rural bunkhouses, or of the physique of decathlon athletes. The failure to fit these listlike paragraphs may indicate a basic limitation of the model: It presumes a well-structured text. When it gets a list of sentences that have no textlike structure, it is unable to adapt its strategies and makes plainly wrong predictions.

Given the overall high recall level, the levels effect (e.g., Kintsch & Keenan, 1973) was much reduced in magnitude. Nevertheless recall decreased significantly as a function of level, $F(1, 19) = 15.36$, $MS_e = .004$, to an asymptote of about 67%. The predictions of the model correspond reasonably well to the observed effect as shown in Figure 1.

Model Predictions: Readability

The correlations between reading time, recall, readability, and a number of selected text properties are shown in Table 5. Although these statistics were computed from only 20 texts and are therefore not very powerful, some interesting trends are demonstrated. Most interesting are the correlations between the dependent variables and statistics derived from the model's simulation of the comprehension process. As predicted, number of inferences and number of reinstatements correlate with readability, with reading time, and to a limited extent, with recall. I and s, the model parameters that ultimately determine reinstatements and (to a lesser degree) inferences are not themselves significantly correlated with the dependent variables. However, as a general trend, larger short-term buffers are associated with better recall, shorter reading times, and greater readability. Larger input chunks also are associated with shorter reading times and better recall. The number of input chunks, which is highly correlated with $I (r = -.81)$ because of the relatively constant length of the paragraphs, reflects the same trends. The number of short-term memory overloads represents a general processing load index, and correlates rather well with the subjective readability ratings and the Flesch (1948) scores for the texts.

Two statistics related to the text base —proposition density and number of arguments—are also of interest. A greater number of arguments lead to longer reading times and have a corresponding effect on readability, but do not appear to affect recall. The correlation between recall and proposition density is spurious: Since the length of the paragraphs was relatively constant, an increased number of words per proposi-

Table 5
Correlations Between Four Sets of Predictor Variables and Three Readability Indices

Predictor variable	Reading time	Recall	Readability
Model statistics			
Inferences	.48*	−.52*	.57*
Reinstatements	.44*	−.34	.61*
Buffer capacity (s)	−.25	.28	−.38
Short-term memory overloads	.28	.00	.18
Input size (I)	−.27	.16	−.28
Cycles	.28	.02	.19
Textbase statistics			
Proposition density	−.03	−.43	.26
Number of arguments	.44*	.00	.36
Text statistics			
Word frequency	−.33	.18	−.31
Sentence length	.09	−.15	.14
Ratings			
Subjective readability	.52*	−.07	.36
Flesch score	−.36	.00	−.22

* Significant at .05 level.

tion means that there were simply fewer propositions to be recalled from that text.

The text statistics of word frequency and sentence length are not significantly correlated with the dependent variables, probably as a result of the limited sample size. The direction of these correlations, however, confirms the expectation that low reading times and high recall should be characteristic of texts with common words and short sentences. Finally, the two rating variables indicate that subjective readability and the Flesch score account for the reading time of a text rather well but that they do not account for how much is retained from the text.

There is no point in trying to develop a new "readability formula" on the basis of this research. Twenty texts is too few on which to base a truly successful formula, and moreover, Kintsch and Vipond (1979) have discussed why another formula would not be desirable. However, an appropriate test of the model would be to attempt to predict the readability of the texts with the statistics that can be derived from the model and the texts. Therefore, stepwise multiple regressions were performed for reading time, recall, and readability using the model, text base, and text statistics as predictors. Variables were entered into the equation corresponding to the percent of variance accounted for by that variable. Given the limited amount of data that constitute these variables, Table 6 shows the construction of the regressions for all predictor variables that maintain a multiple correlation at a significance level of .05 or better. As shown in these tables, the regressions are quite successful, yielding multiple correlations of .83, .85, and .86 for reading time, recall, and readability, respectively.

In addition, the relative power of the predictor variables is much as expected: Reinstatements and inferences are the basic predictors of all three dependent variables. The number of reinstatements does not appear in the reading-time regression because the variance accounted for by reinstatements is almost equal to but slightly less than the variance accounted for by inferences. Hence, the inference factor is entered into the regression equation first, removing from the regression task the variance that can be explained by reinstatements: The partial correlations between reinstatements and reading time with inferences held constant is .002, whereas the simple correlation between reinstatements and reading time is .44 for the set of texts used here.

Note that although word frequency and sentence length are important predictors of reading time and readability, they have little effect on recall. This implies that unfamiliar words and long sentences may slow the reader down, but once the reader moves past the encoding stage, they have little effect on memory. Finally, although Flesch scores and subjective readability correlate

Table 6
Stepwise Regressions for Reading Time, Recall, and Readability

Factor	*R*
Reading time[a]	
Number of inferences	.48
Word frequency	.58
Number of arguments	.64
Number of cycles	.69
Sentence length	.80
Words per proposition	.82
Short-term memory stretches	.83
Input size	.83
Recall[b]	
Number of reinstatements	.52
Words per proposition	.61
Number of inferences	.64
Number of arguments	.69
Number of cycles	.70
Input size	.80
Short-term memory stretches	.81
Buffer size	.84
Sentence length	.85
Readability[c]	
Number of reinstatements	.61
Number of inferences	.71
Word frequency	.74
Input size	.76
Sentence length	.82
Short-term memory stretches	.84
Number of arguments	.85
Buffer size	.85
Words per proposition	.86

[a] $F(8, 11) = 3.15, p = .041$.
[b] $F(9, 10) = 3.01, p = .05$.
[c] $F(9, 10) = 3.02, p = .05$.

with reading time, they are independent of recall and are thus poor predictors of these texts' readability.

Most noteworthy, however, is the interaction between the model and text variables. By themselves, the text variables do not predict a significant amount of the variance for any of the dependent variables. However, when these variables are entered into an analysis with the model variables, the performance of the model variables is greatly enhanced. These analyses thus support one of the major tenets of this article: Readability is an interactive relationship between the properties of a text and the reader who is processing it.

Discussion

Readability is not a property of a text, to be measured by the right kind of formula. The results reported here lend support to an alternative conception of readability, in which the readability of a text is determined by the ways that certain text properties—primarily the arrangement of the propositions in the text base, but also word frequency and sentence length—interact with the reader's processing strategies and resources. The number of inferences that need to be made in the construction of a coherent text base and the number of reinstatements of already processed propositions into working memory appear to be two very important determinants of readability. Their influence depends on the way the text is written as well as on the reader: If a reader is able to process large input chunks and can support a buffer of relatively large capacity, fewer reinstatements and inferences will be necessary. Alternatively, a reader limited in input capacity and buffer size will be plagued by more frequent short-term memory overloads and long-term memory reinstatements. To the extent that these disruptions require cognitive capacity, the reader's ability to generate inferences that bind the text together will also be depressed. The final result is a decrease in reading efficiency.

Different aspects of readability are subject to somewhat different influences: The commonly cited variables of word frequency and sentence length appear to affect reading time but not recall. Once a text segment has been decoded, word frequency no longer matters, presumably because further processing occurs on a conceptual level that is unaffected by these factors. Alternatively, reinstatements and inferences are important for both recall and reading time. They inflate reading time because the normal flow of reading is interrupted by the long-term memory searches. Similarly, when a reader does not reinstate a needed proposition or does not draw a necessary bridging inference, the resulting text base will be incoherent, resulting in major problems during recall.

The multiple correlations between indices of readability and several model-dependent variables were reassuringly high in this experiment, although it would be incorrect to interpret the corresponding regression equations as new readability formulas. Instead, they should be considered as indications that certain model-dependent predictors—primarily reinstatements and inferences—are indeed important determinants of readability.

This model does not at present contain an inference component; it merely predicts when bridging inferences should occur. A properly defined inference component would have to be responsible for a variety of inferences, reflecting the differential effects that inferences can have on the reading process. The present version of the model has no knowledge about the meaning of a text's words and so cannot build coherence graphs aided by the knowledge that as was the case in one of the experimental texts, *acupuncture* is practiced by an *acupuncturist*. These types of inferences, based on simple semantic relations, are probably quite automatic—they should not make significant demands on resources, and consequently they should not interrupt the normal comprehension process. In fact, writing out such well-known information may actually interfere with the readability of a text, rather than improve it (e.g., Keenan & Kintsch, 1974; Meyer, 1975; Shuy & Larkin, 1978).

At the other extreme, there are inferences that without the necessary informa-

tion, simply cannot be made. In one of the experimental paragraphs, it was necessary to know that a certain person was a defendant in a Watergate-related trial. If a reader did not have this information, then the text would seem incoherent, forcing the reader either to continue reading without knowing the appropriate relation between these segments of text or else to attempt to generate an inference that may be inappropriate. In either case, the reader's comprehension of the text will be impaired, as will readability.[5] This seems to have been the case for our subjects: This text had the second longest reading time and the fourth lowest percentage recall of our 20 texts. Hence, the impact of all types of inferences on readability is significant, and the construction of an appropriate knowledge-based inference component must be one of the directions taken in the future development of the model.

The importance of the inference and macrostructure component can be seen elsewhere in the data as well. The recall predictions of this model were only moderately successful. Although the generally significant goodness-of-fit tests can be partially discounted because of the extreme power of the chi-square test, the overall correlation between the observed and predicted recall frequences of $r = .60$ is substantially less than what has been found in other applications of the model. Kintsch and van Dijk (1978) reported correlations between .92 and .96; Spilich et al. (1979) found correlations of .73 and .83 for high- and low-knowledge subjects. Kozminsky, Bourne, and Kintsch's (Note 4) recall predictions for four stock market reports correlated between .81 and .87 with the data. What distinguishes these results from those of the current work is that the models presented by these authors employed a theory combining microprocesses with macroprocesses; here, only the microstructure component was tested. A truly successful model cannot neglect macroprocesses, although this neglect served a useful purpose here, in that we were able to test a component of the model in isolation, with real-world, out-of-context paragraphs for which a theoretical macrostructure could not be clearly defined (as opposed to well-structured research reports, baseball game summaries, and carefully constructed stock market reports). It appears, however, that even with short texts and retention intervals, macroprocesses play a significant role in comprehension.

The ability of the model to predict the levels effect shown in Figure 1 suggests that this effect can be explained to a considerable extent by the microprocesses hypothesized here. For a proposition to be held over for processing on additional cycles, it must be contacted by propositions from these cycles. Such propositions likely contain information that is relevant to the entire text and are therefore judged to be the most important in a text. In this way, the process of argument overlap would guide the construction of the coherence graph so that these global, important propositions are retained for additional processing cycles. It is doubtful that this can be a complete explanation, however. As noted earlier, propositions are often recalled due to their role in the macrostructure of the text. Mandel (1979; cf. Cirilo & Foss, 1980; Just & Carpenter, Note 5) has shown that constituent words of high-level propositions are fixated longer during reading than words corresponding to propositions at the lower levels of the text base hierarchy, suggesting that there is something recognizably important about these propositions even during their initial processing. However, the strongest effect revealed by Mandel's eye movement measurements was a much greater tendency to regress to important than to unimportant words; this is consistent with the rehearsal and reinstatement interpretation of the levels effect offered here. One might ask, therefore, whether macrostructure considerations alone could account for recall data. This does not appear to be the case, however: In all previous applications of the model (Kintsch & van

[5] Note that even if the reader continues through the text and ultimately derives the information that should have been inferred earlier, rereading of the passage would probably be necessary to properly interpret the text in light of this new information. Hence, readability would still be impaired.

Dijk, 1978; Spilich et al., 1979; Kozminsky et al., Note 4), both the micro- and macrocomponents of the theory were necessary to obtain good recall predictions.

The data in this experiment were obtained in a rather artificial laboratory setting, especially with regard to the out-of-context presentation and free recall of the paragraphs. A recent study by Masson (1979) has employed a more natural paradigm and obtained similar results to those presented here. Masson used three of the paragraphs from this experiment embedded in their original context so that the complete texts were about 500 words long. In a comparison of normal reading processes and skimming, his subjects read these texts at paced rates of either 225, 375, or 600 words/min. Subjects were then given a cued recall test that tested only the critical paragraph. The model presented here fit Masson's cued recall data quite well; the sum of the minimum chi-squares for the three paragraphs for the three reading conditions were $\chi^2(68) = 97.43$, 107.57, and 108.18, respectively. For comparison, the minimum chi-squares for these three paragraphs in the present experiment summed to 123.35. Most interestingly, Masson's estimates of the short-term buffer capacity and input size parameters were essentially the same, regardless of the speed with which the subjects read. Only the reproduction probability p changed as a function of reading rate; it decreased from .39 at the normal rate to .17 at the fast skimming rate.

The fact that this model fits Masson's (1979) data as well as our own is important for the interpretation of our results. Our subjects read rather slowly (mean reading speed: 68 words/min). Thus, their data might reflect the effects of conscious memorization more than normal reading processes: Subjects could isolate important propositions and rehearse them, instead of processing the text in more natural ways. However, when subjects are reading from 200 to 600 words/min as in Masson's experiment, such rehearsal appears highly unlikely; nevertheless, the model successfully fits the data from all three reading speeds. We therefore believe that the reading processes were similar in both cases. Furthermore, to the extent that verbatim memorization was important in the present task, it probably served to distort our results from what could be expected from a purely conceptual text analysis.

In a paradigm very different from Masson's (1979), Spilich et al. (1979) compared the fits that their model achieved for high- and low-knowledge readers; like Masson, the differences in recall between these groups was accounted for not by a different I or s value but rather by a different reproduction probability value. We had no way to distinguish our subjects on an a priori basis, but a post hoc attempt was made on the basis of high and low recall levels and reading times. The comparison between high and low recall subjects was done as follows: For each group of 120 subjects, those subjects whose reading times were within one standard deviation of their group's mean were extracted for further analysis; this was done to remove possibly unstable, outlying subjects. These subjects were rank ordered by recall, and the 30 highest and 30 lowest subjects were selected. The propositional recall for these subjects was then collected so that recall patterns for high and low recall subjects were obtained for all 20 texts. These recall patterns were used in the same parameter estimation procedure described earlier to obtain the best-fitting parameter estimates for high and low recall subjects on each text. A similar procedure was carried out to obtain parameter estimates for high and low reading time subjects. The same trends were demonstrated here as by the other researchers cited earlier: The significant differences in recall and reading time manifested themselves not in different buffer sizes and input values but in different values of p. We hasten to emphasize the post hoc nature of this analysis, but it is still worth noting, particularly as a comparison between this class of results and those discussed by Lesgold and Perfetti (1978), in which good and poor readers demonstrated differences in effective short-term memory capacity.

It is unclear why I and s do not change in these applications of the model. We had originally hypothesized that Masson's

results might be characterized by larger input chunks at higher reading rates, and similarly suspected that our post hoc class of "good" readers would be best fit by a model configuration with larger than normal input chunks and short-term buffers. In retrospect, however, it seems reasonable that buffer size be thought of as a property of the cognitive system that would be inherently invariant. The stability of I at high reading speeds might suggest that some selective processing is occurring before a text's propositions are entered into the short-term buffer, either by the selective decoding of only that part of the text that is related to particularly important words or perhaps by the selection of only important propositions for coherence processing.

A reasonably complete model of comprehension must include several components that have been neglected here. As noted earlier, it must deal with macroprocesses. The Kintsch and van Dijk (1978) model is relatively well established in this respect, although we have not yet achieved the same level of formalization as the current study's treatment of microprocesses. Notice, however, that the present model contains some features that interact with the macrostructure of the text. In particular, those propositions chosen for the top of coherence graphs and the propositions selected by the leading edge strategy are almost always relevant to the macrostructure. The leading edge strategy similarly favors propositions at the top of the graph; these propositions are there, of course, because of their local semantic importance to the text. What we refer to as "macroprocesses" are those processes that deal with the global semantic content of one or more propositions, resulting in the transformation of old propositions or the creation of new propositions. Such processes are defined by Kintsch and van Dijk (1978) and van Dijk (1977), but have yet to be incorporated into a working model.

A complete model of comprehension would also contain a semantic parser that would transform literal text into a conceptual text base; we currently bypass that step by handcoding the proposition list. The chunking program reported here is a first step in this direction, in that it processes both the proposition list and the original text. Finally, this model does not explicitly confront the question of how the reader's knowledge sources interact with the text via inference generation. Some preliminary suggestions have been explored in Kintsch and van Dijk (1978) and Kintsch (1979); these ideas are parallel to the discussions of frames (Minsky, 1975), scripts (Schank & Abelson, 1977), and schemata (Rumelhart & Ortony, 1977). The integration of the present model with knowledge-based inferencing systems is a clear and necessary goal, but a more immediate question concerns how these substantial and necessary elaborations will affect the generalizability of the present results.

There is reason to believe that many of the findings relevant to recall and readability would hold up quite well. There would probably be some differences in exactly when and why inferences and reinstatements occur in a more elaborate system, in that the criteria would shift from the disruption of *formal* coherence (argument repetition) to *semantic* coherence (relations among concepts). As in the example cited earlier, a successful reading model should not be significantly disrupted by having to determine that *acupuncture* is performed by an *acupuncturist*. Similarly, little disruption might be expected when, to use the example in the appendix, the occurrence of miracles is discussed in Cycle 1 and a person's sudden recovery from a fatal illness is discussed in Cycle 2. What links these two cycles is really that the incident in Cycle 2 is an example of something that was discussed in Cycle 1. The coherence should focus primarily on that semantic relation, and not on a bishop that was mentioned in both cycles (as the present model does). Hence, the principle of coherence can be retained, and we should expect breakdowns in coherence still to be characteristic of impaired readability.

A final set of considerations is more methodological than theoretical. In the present experiment as well as in previous work, we have relied on the indirect measures of recall and reading time to study comprehension. Since these measures are

explicitly founded in the comprehension theory, their use is legitimate and has proven to be quite informative. As theory construction develops, however, we can make more specific predictions about the comprehension process. Hence, our experimental work should benefit from developments in the theory and from the use of on-line techniques to probe the state of memory processes at particular points in the comprehension process. That is, we must be able to confirm the theory's prediction that a given piece of information is in short-term memory, and to indicate at what point the majority of a subject's processing resources would be allocated to long-term memory retrieval or to inference generation. Such techniques should yield the kind of rich empirical data base that is capable of supporting the more complex theoretical work necessary in this area of investigation.

Reference Notes

1. Bjork, R. Personal communication, 1978.
2. Wickens, T. D. *Parameter estimation in markov chain learning models*. Unpublished master's thesis, Brown University, 1967.
3. Chandler, P. J. *Subroutine STEPIT: An algorithm that finds the values of the parameters which minimize a given continuous function*. Bloomington: Indiana University, Quantum Chemistry Program Exchange, 1965.
4. Kozminsky, E., Bourne, L. E., & Kintsch, W. *Comprehension and analysis of information in text: I. Construction and evaluation of brief texts* (Tech. Rep. 82-ONR). Boulder: University of Colorado, Institute for the Study of Intellectual Behavior, 1979.
5. Just, M. A., & Carpenter, P. A. *Towards a theory of reading comprehension: Models based on eye fixations*. Unpublished manuscript, Carnegie-Mellon University, 1979.

References

Bransford, J. D., & Franks, J. J. The abstraction of linguistic ideas. *Cognitive Psychology*, 1971, *3*, 331–350.

Caccamise, D. J., & Kintsch, W. Recognition of important and unimportant statements from stories. *American Journal of Psychology*, 1978, *91*, 651–657.

Cirilo, R. K., & Foss, D. J. Text structure and reading time for sentences, *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 96–109.

Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948, *32*, 221–233.

Graesser, A. How to catch a fish: The representation of memory of common procedures. *Discourse Processes*, 1978, *1*, 72–89.

Haviland, S. E., & Clark, H. H. What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 1974, *13*, 515–521.

Johnson, R. E. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 12–20.

Keenan, J. M., & Kintsch, W. The identification of explicitly and implicitly presented information. In W. Kintsch (Ed.), *The representation of meaning in memory*. Hillsdale, N.J.: Erlbaum, 1974.

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

Kintsch, W. On modeling comprehension. *Educational Psychologist*, 1979, *14*, 3–14.

Kintsch, W., & Keenan, J. M. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 1973, *5*, 257–274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. Comprehension and recall of texts as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 1975, *14*, 196–214.

Kintsch, W., & Polson, P. G. On nominal and functional serial position curves: Implications for short-term memory models? *Psychological Review*, 1979, *4*, 407–413.

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

Kintsch, W., & Vipond, D. Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Ed.), *Perspectives on memory research*. Hillsdale, N.J.: Erlbaum, 1979.

Kučera, H., & Francis, W. N. *Computational analysis of present-day American English*. Providence, R.I.: Brown University Press, 1967.

Lesgold, A. M., & Perfetti, C. A. Interactive processes in reading comprehension. *Discourse Processes*, 1978, *1*, 323–336.

Mandel, T. S. Eye movement research on the propositional structure of short texts. *Behavior Research Methods and Instrumentation*, 1979, *11*, 180–187.

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

Masson, M. E. J. *Cognitive processes in skimming stories*. Unpublished doctoral dissertation, University of Colorado, 1979.

McKoon, G. Organization of information in text memory. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 247–260.

McKoon, G., & Keenan, J. M. Response latencies to explicit and implicit statements as a function of the delay between reading and test. In W. Kintsch (Ed.), *The representation of meaning in memory*. Hillsdale, N.J.: Erlbaum, 1974.

Meyer, B. J. F. *The organization of prose and its effect upon memory*. Amsterdam: North-Holland, 1975.

Meyer, B. J. F., & McConkie, G. W. What is recalled after hearing a passage? *Journal of Educational Psychology*, 1973, *65*, 109–117.

Minsky, M. A. A framework for representing knowl-

edge. In P. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, 1975.

Rumelhart, D. E., & Ortony, A. The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge*. Hillsdale, N.J.: Erlbaum, 1977.

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

Schank, R. C., & Abelson, R. P. *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Erlbaum, 1977.

Shuy, R. W., & Larkin, D. L. Linguistic considerations in the simplification/clarification of insurance policy language. *Discourse Processes*, 1978, *1*, 305–321.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 1979, *18*, 275–290.

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

Turner, A., & Greene, E. Construction and use of a propositional text base. JSAS *Catalog of Selected Documents in Psychology*, 1978, *3*, 58. (Ms. No. 1713)

van Dijk, T. A. Semantic macro-structures and knowledge frames in discourse comprehension. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension*. Hillsdale, N.J.: Erlbaum, 1977.

Waters, H. S. Superordinate–subordinate structure in semantic memory: The roles of comprehension and retrieval processes. *Journal of Verbal Learning and Verbal Behavior*, 1978, *17*, 587–598.

Appendix

Computer Trace of the Coherence Graph Generation for the "Saint" Sample Paragraph: Minimum Words for Input $I = 19$; STM buffer size $s = 2$

Input for Cycle 1:

(P1 (REQUEST P2 P8))
(P2 (CANONIZE P3))
(P3 (ISA JOHN-NEWMANN FRONTIER-PRIEST))
(P4 (ISA JOHN-NEWMANN BISHOP))
(P5 (LOC:IN P4 PHILADELPHIA))
(P6 (TIM:IN P4 19TH-CENTURY))
(P7 (TWO MIRACLES))
(P8 (ATTRIBUTED P7 JOHN-NEWMANN))
(P9 (TIME:IN P8 THIS-CENTURY))

Build a graph with Proposition P1 at Level 1.
Put P2 at Level 2, pointed to by P1.
Put P8 at Level 2, pionted to by P1.
Put P9 at Level 2, pointed to by P1.
Put P3 at Level 3, pointed to by P8.
Put P4 at Level 3, pointed to by P8.
Put P7 at Level 3, pointed to by P8.
Put P5 at Level 4, pointed to by P4.
Put P6 at Level 4, pointed to by P4.

Apply the leading edge strategy.
This is Cycle 1: STM is expanded to 3 slots for this cycle only.
Retain Proposition P1 at Level 1.
  Proposition P2 is embedded in P1: retain Proposition P2 at Level 2.
  Proposition P3 is embedded in P2: retain Proposition P3 at Level 3.
  Proposition P8 is embedded in P1: retain Proposition P8 at Level 2.
  Proposition P7 is embedded in P8: retain Proposition P7 at Level 3.
The buffer is currently overloaded by more than one proposition. The embedded Proposition P3 must be deleted to remove the overload. At the end of Cycle 1, STM contains:
Level 1:
  P1 points to (P2 P8)
  Level 2:
    P2 points to nothing.
    P8 points to P7
    Level 3:
      P7 points to nothing.

Input for Cycle 2:

(P10 (TIME:IN P11 1923))
(P11 (DYING EVA-BENASSI PERITONITIS))
(P12 (DRAMATICALLY P13))
(P13 (RECOVERED EVA-BENASSI))
(P14 (AFTER P15 P13))
(P15 (PRAYED NURSE BISHOP))

Add the Cycle 2 propositions to memory.
Propositions (P10 P11 P12 P13 P14 P15) cannot be added to STM: Search LTM for a new starting proposition.
The LTM search succeeded: P4 can be reinstated via P15.
Bump P4's cycle counter.
Nothing from INPUTSET was placed, but the LTM search succeeded: P4 was found.
This counts as a reinstatement search.

Build a graph with Proposition P13 at Level 1.
Put P11 at Level 2, pointed to by P13.
Put P12 at Level 2, pointed to by P13.
Put P14 at Level 2, pointed to by P13.
Put P15 at Level 3, pointed to by P14.
Put P10 at Level 3, pointed to by P11.
Put P4 at Level 4, pointed to by P15.

Apply the leading edge strategy.
Retain Proposition P13 at Level 1.
Retain Proposition P14 at Level 2.
  Proposition P15 is embedded in P14: retain Proposition P15 at Level 3.
The buffer is stretched to one more than its set size = 3. At the end of Cycle 2, STM contains:
Level 1:
  P13 points to P14
  Level 2:
    P14 points to P15
    Level 3:
      P15 points to nothing.

Input for Cycle 3:

(P16 (TIME:IN P17 1949))
(P17 (HOSPITALIZED KENT-LENAHAN P18 P20 P21))
(P18 (FRACTURES SKULL KENT/LENAHAN))
(P19 (TWO P18))
(P20 (SMASHED BONES KENT/LENAHAN))
(P21 (PIERCED LUNG KENT-LENAHAN))
(P22 (AFTER P17 ACCIDENT))
(P23 (TRAFFIC ACCIDENT))
(P24 (ROSE KENT-LENAHAN DEATHBED))
(P25 (RESUMED KENT-LENAHAN P26))
(P26 (NORMAL LIFE))

Add the Cycle 3 propositions to memory.
Propositions (P16 P17 P18 P19 P20 P21 P22 P23 P24 P25 P26) cannot be added to STM: search LTM for a new starting proposition.
Nothing from INPUTSET was placed, and the LTM search failed.
This counts as an inference.
Build a graph with Proposition P18 at Level 1.
Put P17 at Level 2, pointed to by P18.
Put P19 at Level 2, pointed to by P18.
Put P20 at Level 2, pointed to by P18.
Put P21 at Level 2, pointed to by P18.
Put P24 at Level 2, pointed to by P18.
Put P25 at Level 2, pointed to by P18.
Put P26 at Level 3, pointed to by P25.
Put P16 at Level 3, pointed to by P17.
Put P22 at Level 3, pointed to by P17.
Put P23 at Level 4, pointed to by P22.
Apply the leading edge strategy.
Retain Proposition P18 at Level 1.
Retain Proposition P25 at Level 2.
  Proposition P26 is embedded in P25: retain Proposition P26 at Level 3. The buffer is stretched to one more than its set size = 3.
  At the end of Cycle 3, STM contains:
Level 1:
  P18 points to P25
  Level 2:
    P25 points to P26
    Level 3:
      P26 points to nothing.

Input for Cycle 4:
(P27 (AFTER P25 P28))
(P28 (PRAYED MOTHER JOHN-NEWMANN))
(P29 (ARDENTLY P28))

Add the Cycle 4 propositions to memory.
Put P27 at Level 3, pointed to by P25.
Put P28 at Level 4, pointed to by P27.
Put P29 at Level 4, pointed to by P27.
Apply the leading edge strategy.
Retain Proposition P25 at Level 1.
Retain Proposition P27 at Level 2.
  Proposition P28 is embedded in P27: retain Proposition P28 at Level 3.
The buffer is stretched to one more than its set size = 3.
At the end of Cycle 4, STM contains:
Level 1:
  P25 points to P27
  Level 2:
    P27 points to P28
    Level 3:
      P28 points to nothing.

End of file found.
The final LTM graph contains 2 subgraphs.
This run required 4 input cycles, 1 reinstatement, and 1 inference.
STM was stretched 4 times and overloaded 1 time.
13 propositions were given extra processing:
  Proposition P1: 1 cycle.
  Proposition P2: 1 cycle.
  Proposition P4: 1 cycle.
  Proposition P7: 1 cycle.
  Proposition P8: 1 cycle.
  Proposition P13: 1 cycle.
  Proposition P14: 1 cycle.
  Proposition P15: 1 cycle.
  Proposition P18: 1 cycle.
  Proposition P25: 2 cycles.
  Proposition P26: 1 cycle.
  Proposition P27: 1 cycle.
  Proposition P28: 1 cycle.

Received October 15, 1979
Revision received February 12, 1980 ■