Models for the Design of Instructional Text

Thomas M. Duffy; Lorraine Higgins; Brad Mehlenbacher; Cynthia Cochran; David Wallace; Charles Hill; Diane Haugen; Margaret McCaffrey; Rebecca Burnett; Sarah Sloane; Suzanne Smith

Reading Research Quarterly, Vol. 24, No. 4. (Autumn, 1989), pp. 434-457.

Stable URL:
http://links.jstor.org/sici?sici=0034-0553%28198923%2924%3A4%3C434%3AMFTDOI%3E2.0.CO%3B2-6

Reading Research Quarterly is currently published by International Reading Association.

Thomas M. Duffy
Indiana University

Lorraine Higgins
Brad Mehlenbacher
Cynthia Cochran
David Wallace
Charles Hill
Diane Haugen
Margaret McCaffrey
Rebecca Burnett
Sarah Sloane
Suzanne Smith
Carnegie Mellon University

Models for the design of instructional text

IN ORDER to identify guidelines for writing textbooks that better facilitate students' learning, a number of researchers have analyzed the revision strategies of experts and the effects of their revisions. The present study partially replicates and extends a study originally conducted by Graves et al. (1988) comparing a segment from a high school history text and revisions of that text by text linguists, composition teachers, and Time-Life writers. In revising the text, the Time-Life writers emphasized making the text more interesting, whereas the other two groups of experts focused on clarifying the structure and providing cohesion. In the current study, the performance of 11th-grade students on free-recall and short-answer tests as well as their ratings of interest and ease of learning all indicate that the revision by the composition teachers was most effective. This result confirms prior work by Britton, Van Dusen, Gulgoz, & Glynn (1989), but fails to confirm the original study by Graves et al., in which superior recall was found for the revision by the Time-Life writers. The authors further analyze the information recalled by students in relation to the hierarchical structure of the text in an attempt to understand the experts' specific revision techniques. The authors suggest that presenting textbook writers with a series of text-independent guidelines may be less effective than alternative strategies such as protocol-aided revision, in which the text is revised based on comments made by readers as they try to understand it.

Modèles de construction de textes didactiques

POUR ISOLER des règles de production de textes dans les manuels scolaires qui rendent l'apprentissage plus facile, certains chercheurs ont analysé les stratégies de révision de textes utilisées par des experts et évalué les effets de ces révisions. La présente étude reprend et prolonge une étude semblable menée par Graves et al. (1988) comparant les révisions effectuées par des linguistes, des professeurs de français et des auteurs du *Time-Life* sur un extrait de texte d'un manuel d'histoire du niveau collégial. Les changements apportés par les auteurs du *Time-Life* mettaient l'accent sur l'intérêt du texte alors que les deux autres groupes se centraient sur la structure et la cohérence textuelles. Les résultats obtenus, après lecture des textes revisés, par des élèves de onzième année, à des mesures de rappel, de compréhension de textes de même qu'à l'évaluation de l'intérêt et de la difficulté de lecture des textes, montraient que ce sont les révisions effectuées par les professeurs de français qui ont été les plus efficaces. Ces résultats confirmaient ceux obtenus par Britton, Van Dusen, Gulgoz, et Glynn (1989), mais contredisaient ceux obtenus dans la précédente étude (Graves et al., 1988) alors que ce sont les révisions des auteurs du *Time-Life* qui s'étaient révélées les plus efficaces. Les auteurs ont plus particulièrement examiné, dans les rappels des élèves, la macrostrucuture sémantique pour essayer de mieux cerner les effets des techniques spécifiques de révision des experts. Les résultats semblent indiquer que de présenter aux auteurs de manuels scolaires une série de directives générales, indépendantes du texte, produit moins d'effets qu'une stratégie de réviser avec l'aide de protocoles, consistant à intégrer les commentaires faits par des lecteurs placés en situation de compréhension du texte.

Modelos para el diseño de libros de texto

PARA PODER identificar lineamientos para escribir libros de texto que faciliten más el aprendizaje de los estudiantes, algunos investigadores han analizado las estrategias de revisión de los expertos en el área y los efectos de sus revisiones. Este estudio replica y extiende otro llevado a cabo originalmente por Graves et al. (1988) que comparaba un segmento de un libro de historia de nivel secundario y las revisiones hechas al texto por lingüístas, maestros de composición y escritores de la editorial *Time-Life*. Al revisar el texto, los escritores de *Time-Life* enfatizaron el hacer el texto más interesante, mientras que los otros dos grupos de expertos enfocaron sus esfuerzos en clarificar la estructura del texto y en proporcionar cohesión. El desempeño de los estudiantes de onceavo grado en recuerdo libre y pruebas de respuesta corta así como sus juicios respecto al interés del texto y la facilidad de aprenderlo, indicaron que la revisión hecha por los maestros de composición fue la más efectiva. Este resultado confirma el trabajo previo de Britton, Van Dusen, Gulgoz y Glynn (1989), pero falla en confirmar el estudio original (Graves et al., 1988), en que se encontró que el recuerdo de la revisión hecha por los escritores de *Time-Life* fue superior. Los autores además analizaron más profundamente la información recordada por los estudiantes en relación a la estructura jerárquica del texto en un intento por comprender mejor las técnicas específicas de revisión usadas por los expertos. Se sugiere que presentar a los escritores de libros de texto con una serie de lineamientos no relacionados con el texto puede ser menos efectivo que proporcionar estrategias alternativas tales como revisión con ayuda de protocolos, en la que el texto es revisado basándose en comentarios hechos por los lectores mientras trataban de entenderlos.

Modelle zum Entwurf von Lehrtexten

UM RICHTLINIEN zum Schreiben von Lehrbüchern, die das Erlernen seitens der Schüler weiterhin erleichtern, zu erkennen, befaßten sich einige Forscher mit den Korrekturstrategien von Experten und den Auswirkungen der Korrekturen. In der vorliegenden Arbeit wurde eine Studie, die ursprünglich von Graves und Mitarbeitern 1988 durchgeführt worden war, wiederholt und weiter ausgebaut. Dabei wurde ein Ausschnitt, der aus einem *High School* Geschichtsbuch stammte, mit den Korrekturen, die von Textlinguisten, Lehrern, die Schulaufsatztechniken unterrichten, und Redakteuren der Zeitschrift *Time-Life* vorgenommen worden waren, verglichen. Bei der Korrektur des Textes betonten die *Time-Life* Redakteure, daß der Text interessanter gemacht werden sollte, während die beiden anderen Expertengruppen sich darauf konzentrierten, die Struktur zu vereinfachen und den Zusammenhang herzustellen. Die Leistung von Schülern der 11. Klasse, sowohl bei einer freien Nacherzählung und Testen, bei denen kurze Antworten zu geben waren, als auch bei deren Einstufung des Textinteresses und der Natürlichkeit des Lernprozesses, deuten allesamt darauf hin, daß die Korrektur der Lehrer am wirksamsten war. Dieses Ergebnis bestätigt die Ergebnisse von Britton, Van Dusen, Gulgoz und Glynn (1989), aber sie bestätigt nicht die Ergebnisse der Originalstudie (Graves und Mitarbeiter, 1988), bei der die Korrektur seitens der Redakteure bessere Daten bei der Nacherzählung ergab. Des weiteren untersuchten die Verfasser dieser Arbeit auch die Informationen, die von Schülern beim Nacherzählen vermittelt wurde, im Hinblick auf die hierarchische Struktur des Textes, um die speziellen Korrekturtechniken der Experten verstehen zu können. Die Verfasser schlagen daher vor, daß es wahrscheinlich nicht so wirksam ist, wenn Lehrbuchautoren eine Reihe von textunabhängigen Richtlinien erhalten; alternative Strategien, wie z.B. formative Bewertungen, bei denen der Text anhand der Kommentare von Lesern, die während des Lesens versuchen, den Text zu verstehen, korrigiert wurde, dürften jedoch wirksamer sien.

The textbook is at the core of the curriculum in most schools, and it is typically the basis for creating knowledge that is shared between teacher and students. In short, the textbook is a major determinant of what is taught and how it is taught. Given the central importance of the textbook in our educational system, it is not surprising that both researchers and administrators have paid considerable attention to its design and content. The primary goal of both groups is to make the information in the textbook easier to understand and learn.

During the 1970s researchers generally accepted the core instructional text as given, and focused instead on providing adjunct text, questions, or exercises that would provide the student with an organizational structure for the text, direct attention to the key information in the text, and promote the appropriate level and type of processing of the text. The variety of adjunct aids studied by these researchers included inserted questions, advance organizers, review questions, summaries, applications, and problem-solving exercises (Anderson & Biddle, 1975; Mayer, 1979, 1984; Rickards, 1979). Publishers in turn incorporated adjunct aids into textbooks on a grand scale. Indeed, in texts published today for upper elementary students, one frequently finds more space devoted to the adjunct aids than to the instructional text and instructional graphics. The material to be learned is often buried in learning aids.

An alternative to developing adjunct aids is to reconsider the principles followed in designing the actual text, the conveyer of the information to be learned. How should text be written so that it best facilitates learning? That is, what rhetorical principles should be followed in designing and writing instructional text? Until recently, most research addressing this question dealt with the development and application of readability formulas, which were originally designed as a means of assessing the relative *difficulty* of a text (based on the length of the words

and sentences, predictability, etc.). As with adjunct learning aids, publishers and selection committees have made widespread use of readability formulas and guidelines for readable writing. Today, textbooks are, and must be, written to specific readability levels. Such procedures have been universally adopted despite overwhelming evidence against using readability formulas as a guide for writing or as a method for matching a text to a reader (Davison & Green, 1988; Duffy, 1985; Duffy & Kabance, 1982; Kniffin et al., 1980).

One point that must be considered in designing texts is what kind of learning we want to facilitate. If we want the student simply to remember the main ideas, then, as research by Reder suggests (Reder, 1985; Reder & Anderson, 1980), we should present the main ideas in a brief summary. Of course, the communication goal of textbooks extends beyond the simple recall of main ideas. Our goal is to provide a well-organized network of information so that the student can understand the interrelations between concepts and apply them. Text written to comply with readability guidelines does not necessarily provide this kind of structuring of information for the student.

Recent research on text comprehension has suggested principles for text design that extend well beyond readability. We generally recognize that readers use their knowledge of the subject matter and of various text structures to guide their reading and learning. Readers attempt to identify the text structure imposed by the author, and ideally they form a representation in memory parallel to that structure (Meyer, 1984). Readers who are able to identify and use the author's structure are more likely to remember what the author considered to be the key concepts and to organize the information in memory in a way that is consistent with the author's organization. It seems logical that such processes would be facilitated by designing texts that help students to relate their prior knowledge to the text, and that make the structure and organization of the text easier to recognize.

Meyer (Meyer, 1984; Meyer, Brandt, & Bluth, 1980; Meyer & Rice, 1981) found that clearly signaling the structure of a text facilitates recall of the text, especially recall of the main ideas or top-level structure. She and her associates added phrases such as *"A problem...,"* *"The solution to this problem...,"* or *"The trouble is..."* at the beginning of key paragraphs to signal a problem/solution structure and to draw the reader's attention to the main ideas in that structure. Similarly, for a text constrasting opposing views on a controversial topic, they added phrases such as *"Not everyone approved...,"* *"[Certain people] were in favor...,"* and *"Various groups of people opposed..."* to signal the comparison/contrast structure and to emphasize the main ideas in that structure. Meyer and associates also experimented with adding details such as specific names and dates, which are typically at a lower level and do not signal the central structure of a text. They found that addition of this kind of lower-level detail reduced recall of the main ideas, perhaps by distracting the reader's attention from the central structure. This and other research suggests that learning will be facilitated by texts designed to highlight text structure, point the reader to the main ideas, and clearly relate supporting detail to those main ideas.

Although findings from comprehension research have important implications for text design, such research is generally based on texts that are constructed by the experimenter to manipulate a limited number of variables that are hypothesized to affect comprehension. An alternative strategy to identify rhetorical guidelines for instructional texts is to analyze the revision strategy of expert text revisers (Britton, 1986; Duffy, Curran, & Sass, 1983; Okamoto, Calfee, Varghese, & Chambliss, 1987). The basic strategy is to ask various kinds of experts to revise a text, and then evaluate the expert revisions to determine whether those revisions in fact lead to better learning. The researcher then analyzes the texts that do facilitate the desired learning, and if possible also consults the revisers, in order to determine what strategies and rhetorical principles led to the positive effect. This approach permits the researcher to examine the often tacit knowledge of a wide variety of experts, and may help identify variables not

previously considered important. This strategy also assures that the researcher's conclusions will be based on a text and a revision that are ecologically valid.

In one such study, Graves et al. (1988; also reported in Graves & Slater, 1986; Graves et al., 1985) asked pairs of experts from three different backgrounds to revise two 400-word, 11th-grade history texts. The experts included one pair of composition teachers, one pair who worked for Time-Life Books (as an editor and a writer, respectively), and one pair of text linguists. Each pair was asked to work together and to use whatever knowledge they had to make the texts more comprehensible. Each pair of experts also revised the texts a second time, after they had learned how recall performance on their revision compared with recall of the original. On both the first and the second set of revisions, students who read the Time-Life version achieved higher recall scores than students who read the original or either of the other revisions. Furthermore, editors from both the *Christian Science Monitor* and the *American Educator* chose the Time-Life revision as the most effective rewrite, and a member of the Editorial Board of *USA Today* "praised the rewrite lavishly" and described it as "far superior" (Graves et al., 1985, p. 31). Thus, both objective performance and subjective expert judgment identified the Time-Life writing strategy as very effective.

Graves et al. (1988) next examined the revisions and interviewed the writers in order to identify the rhetorical strategies employed. Both the composition teachers and the text linguists reported that they had focused on clarifying the content by highlighting the main ideas, adding cohesive ties within the text, providing background information to help relate the new information to the students' prior knowledge, and deleting irrelevant information. Overall, the emphasis of both of these pairs of revisers was on clarifying the structure of the text and providing cohesion. In contrast, the Time-Life revisers focused on the content. They found the text "dry" and "woefully inadequate" (p. 248). In their revision they attempted to build a sense of drama and to enrich the content with "vivid anecdotes and details that remind us that PEOPLE, not events, make history" (p. 248). Whereas the other revisions were approximately the same length as the original passage, the Time-Life revisers increased passage length by over 80 percent. The increase was due primarily to the addition of anecdotes. They also substituted strong, vivid verbs whenever possible and added colloquialisms and metaphors (e.g., "the Vietnam Bomb Explodes" for the rapid increase in American forces in Vietnam in 1964).

In sum, the "winning" approach was not one that strove for coherence or for highlighting main ideas, but rather one that emphasized vivid language and personal anecdotes. We suspect that these findings are consistent with expectations among the general public about what will "work" in getting children to attend to and learn from text. Although many serious writers and scholars may abhor the writing style of the weekly newsmagazine, we believe most people would accept that this style is generally effective for getting people to pay attention to information—or, at least, more effective than the traditional style of the textbook.

However, although probably consistent with popular belief, the Graves et al. findings are inconsistent with research in text comprehension. The anecdotes added by the Time-Life revisers significantly increased the length of the text, yet they were irrelevant to the main ideas of the text. Reder and Anderson (1980) found that such irrelevant elaborations actually detract from learning the main ideas. Furthermore, the personalized anecdotes added by the Time-Life revisers tended to emphasize ideas that were less important—that is, propositions that were at a low level in the structure of the text. As mentioned earlier, Meyer (1984) found that providing a great deal of detail about ideas that were at a low level in the text structure detracted from recall of the ideas at higher levels. Moreover, although research suggesting that vividness and concreteness facilitate recall is by now well established (e.g., Paivio, 1971), these variables facilitate recall of the particular text units to which they are applied. If the increased vividness and concreteness of the Time-Life text

were restricted mainly to the anecdotes, then the tendency would be even greater to recall the low-level propositions (to which the anecdotes related) better than the main ideas.

Britton et al. (1989) attempted to replicate the Graves et al. findings, but found that the Time-Life version was recalled no better than the original. Instead, the revision by the composition teachers resulted in superior performance both on free recall and on a short-answer test administered immediately after reading. The same result was found for both measures on delayed tests given 1 day later. However, there were no differences between text versions on the recognition test. Because the information was recognized equally well by all students, we infer that there was no difference in how much information was stored after reading the various passages. Rather, it would appear that the revision developed by the composition teachers facilitated retrieval of the information stored.

Current study

Because the findings of the Graves et al. study (1988, unless otherwise noted) are inconsistent with much of the research on text comprehension, because they have so far failed to replicate, and because there were a number of methodological problems that might have accounted for their surprising results, we attempted a partial replication and extension of their study, with some differences in design. First, we question the conclusion of Graves et al. that interest may be more important than text structure in promoting learning, because the greater recall of the Time-Life version was due to the increased text interest provided by the vivid language, the personal scenarios, and the human drama. Such a conclusion is unwarranted because there are no data to suggest that the text was indeed more interesting to the students. Although the Time-Life editors added vivid text and presented personalized scenarios, we do not know whether the *students* found the resulting text more interesting, or whether they found it easier to *learn*. We do know that the *authors* felt it was more interesting. And, as reported elsewhere (Graves et al., 1985), newspaper editors also thought it was more interesting than the original (although they did not get to see the other revisions). However, the students are the learners of interest, and research has shown that the judgment of the expert does not always coincide with that of the novice. In fact, the inability of the expert to see the text through the eyes of the learner may lie at the heart of the difficulty in producing effective textbooks. Therefore, in the present study, after students read each text we asked them to rate how interesting it was and how easy it was to learn, relative to the textbooks they were currently using.

Second, it is important to eliminate the possibility that students recalled more from the Time-Life version simply because the material that was *added* by the Time-Life team was easier to recall. Graves et al. attempted to control for this possible confound by scoring the recalls against the idea units in the original passage. They reported that the students who read the Time-Life version of the text scored higher on this analysis than students who read any other version, including the original text itself. More of the original idea units were recalled from the Time-Life version even though it had the least content in common with the original; it is the version that most dramatically changed the content of the text. We find these data problematic, and we question the logic of scoring literal recall of the revisions according to the content of the original (presumably less effective) text. Although the use of a common scoring grid is not unusual in text comprehension research, usually the researcher first decides on the content that is to be learned, and then embeds that content in different contexts (and the to-be-learned content is what is scored). Instead, in the Graves et al. study, the revisions were scored against *all* of the idea units in the original text—whether or not those idea units were relevant to the content to be learned. Such a scoring procedure would be justified if the designers of the alternative materials had been informed that the criterion for successful redesign would be readers' recall of all idea units in the original passage. Because they were not so informed, some of the revisers may have believed that the goal of the revision was for readers to be able to recall and apply the main ideas. Thus they might, for example, have

left out details from the original text that seemed less important or less relevant to learning the main ideas. Conversely, students who read the Time-Life version may have recalled more of the detail units from the original text, some of which may have been irrelevant to the main ideas. Therefore, we find the data from this analysis difficult to interpret.

In the current study, we sought to determine exactly which parts of the text were recalled better as a function of each type of revision. We hypothesized that the Time-Life version should facilitate recall of low-level propositions in the text, whereas the other revisions, which emphasized coherence, should facilitate the recall of the main ideas and other top-level information. Such findings would be in keeping with the research discussed earlier (e.g., Meyer, 1984; Reder & Anderson, 1980) and with a common-sense notion of the effects of coherence. To test these hypotheses, we scored students' recall of each revision in relation to the position of each idea in the hierarchical structure of the text read.

Third, we attempted to address several methodological flaws in the Graves et al. study, which were identified by Britton et al. (1989). The first problem is that the three revisions were tested with subjects from three different populations. Essentially, each version was evaluated separately against the original in a series of substudies. The student populations differed significantly in terms of both geographic location and socioeconomic status (SES). Graves et al. acknowledged the problem but suggested that they knew of no reason why there should have been a student-by-treatment interaction. However, findings from any number of studies of learning strategies and metacognitive processing (e.g., Brown & Day, 1983; Brown & Smiley, 1978; Wagner & Sternberg, 1987) do suggest that SES differences can enhance or reduce the effects of such text manipulations. That such a confound may in fact have affected the Graves et al. results is suggested by the large difference in mean recall—almost half of a standard deviation—between the three control groups (Graves et al., 1988, Experiment 1).

The second methodological problem was that different experimenters tested subjects and scored the data for the three substudies: The Time-Life revisers tested the effectiveness of their revisions, the text linguists tested theirs, and the composition instructors tested theirs. There is sufficient evidence on experimenter bias to recognize the problems with this procedure. Moreover, there was apparently little coordination between the different experimenters, and thus there may have been undocumented variations in procedure at each site. One such variation was reported: In the first experiment, one of the sites implemented a between-groups experimental design rather than the within-group design used at the other two sites.

A third methodological problem was that the design employed by two groups in Experiment 1 and by all groups in Experiment 2 of Graves et al. was fully within-subject. That is, all subjects received both the original and a revised version of that text. Although the order of presentation was counterbalanced, it is nonetheless difficult to interpret the potential differential carryover from reading two versions of a text on the same topic. As Britton et al. (1989) suggest, "this re-reading might interact with the versions in unknown ways that no counterbalancing scheme could be relied upon to remove."

Finally, we attempted to avoid one possible problem with the Britton et al. (1989) replication. The conflicting findings reported by Britton et al. could have been due to testing the materials with college freshmen, rather than with 11th-grade students, as Graves et al. did. Although Britton et al. reported that the average grade reading level of the participants was 9.5, we are concerned about the students' familiarity with text structures. That is, the college students, because they were the kind of students who had been selected to enter college and because they had 2 more years of school experience, may have been more prepared to identify the structure used by the author and to organize their recall accordingly. Thus, they would have benefited most from the revision by the composition teachers, which emphasized structure and coherence (Meyer, Brandt, & Bluth, 1980;

Meyer & Rice, 1981). Eleventh-grade students, on the other hand, would be less prepared to identify and use the structure of the text; thus, the vividness of the Time-Life approach may well be the most appropriate rhetorical strategy for helping them to maintain attention to the text. In the present study, we attempted to replicate the Graves et al. study more closely by testing the passages using age-appropriate students.

Method

Subjects

The participants were 11th-grade students at a suburban high school. Testing was done in the class, and all students were offered the opportunity to participate. Participation was voluntary, and we offered 5 dollars to the person who obtained the highest recall score in each class. A total of 268 students in 16 classes (over 95% of the population) volunteered.

Students were randomly assigned by class to either the free-recall or the short-answer testing condition. We later discarded the data from 19 students in the free-recall condition who did not write anything on the answer sheet or who wrote irrelevant information (e.g., song lyrics). Thus, data are reported for a total of 122 students in the free-recall condition and 127 students in the short-answer test condition.

Students were also assigned to high or low reading ability groups based on their performance on the reading comprehension portion of the Iowa Test of Basic Skills, administered by the school during the school year. Scores on the test ranged from 40 to 208, with a median of 134; students were divided into the two groups at the median. Reading scores were not available for 10 students in the free-recall condition and 17 students in the short-answer condition; therefore, data from these students were discarded from any analyses involving reading ability, but were included in other analyses.

Materials

Only the Korea passage from Graves et al. was used in this experiment. The results were very consistent across the two passages (Korea and Vietnam) in both the Graves et al. and the Britton et al. studies: Britton et al. obtained a correlation of .92 between mean performance on the two passage topics. Therefore, we felt it was unnecessary to replicate across passages in the present experiment. The Korea passage was selected because it occurred further back in history; hence, we felt the content would be less familiar to the students.

In the Graves et al. study, the original passage was revised twice by the three pairs of revisers: two composition teachers, two text linguists, and the Time-Life writer and editor. The second revision was made after each pair of revisers had learned how well students recalled their version relative to the original. Only the second revision by each pair of revisers was used in the present experiment. A complete description of the revision strategies is presented in Graves et al. (1988). In addition, Britton et al. (1989) provide a description of the physical characteristics of the passages. The length and Kincaid Flesch readability scores (Kincaid, Fishburne, Rogers, & Chissom, 1978) of the original and each revision are presented in Table 1.

Measures

As mentioned above, students were assigned by class to either the free-recall or the short-answer testing condition. For the free-recall test, students simply wrote down as much as they could remember from the text.

The short-answer test consisted of 9 literal comprehension questions. The questions were virtually identical for the four conditions, although wording was changed slightly to reflect the vocabulary used in each revision. The questions were designed to assess the learning of factual information that was presented in all four versions of the text. Sample questions and the relevant portion of each text are presented in the Appendix.

We also solicited interest ratings. We asked the students to rate how easy it was to learn the information in the text, and how enjoyable the text was. In making the ratings, they were asked

Table 1 Descriptive date on four text versions

Measure	Text version			
	Original	Text linguists	Composition teachers	Time-Life
Number of words	412	441	403	730
Number of sentences	23	21	28	35
Readability[a]	11.1	13.2	9.4	12.4

[a]Kincaid Flesch readability scores (Kincaid, Fishburne, Rogers, & Chissom, 1975), as reported by Van Dusen, Britton, and Glynn (1987). Used by permission.

to judge the text relative to the history text they were currently using (or had last used). The rating was done on a five-point Likert scale; the midpoint on the scale was "the same" (in ease or enjoyment, respectively) as the text they were using in school.

Procedure

A packet of all experimental materials for the assigned text version was distributed to each student in a class. We told the students that we were evaluating different ways of writing textbooks, and that their job was to study and learn a page or two from a history text. They were told that they would have 10 minutes to read and study the passage; if they finished early, they should reread and study it just as they would for a test. They were also given instructions for whichever test was assigned to that class ("Write down as much as you can remember" or "Answer some short-answer questions"). Because we observed that all students finished reading and reviewing within 6 minutes, we reduced the time allowed to 8 minutes about halfway through the experiment. Even after this change, all students completely finished reading and reviewing within the time allowed.

At the end of the allotted study time, we asked the students to tear off the passage sheets and place them on the floor. The next sheet was a page of 50 multiplication problems. The participants were told that we needed a measure of their ability, so we wanted them to answer as many of the math problems as they could in 3 minutes. We emphasized that the more problems they answered, the higher their ability score would be. This intervening test was used simply as a means of removing the text information from immediate working memory. We wanted the recall test to reflect storage in long-term memory.

After the 3 minutes, we asked the students to tear off the math sheet and place it on the floor. The next sheet was the recall test. We gave instructions for either the free-recall or the short-answer test (depending on the class). For both, we emphasized that students should guess if they were not sure and that unlimited time was allowed for recall.

Finally, the students turned to the last sheet, which presented the two questions asking them to rate the passage read for ease of learning and for interest.

Text analysis

The text for each of the four passages was analyzed into propositional units similar to those developed by Kintsch & van Dijk (1978) to be used as the basis for scoring free recall. These units consisted of verb phrases (verbs with their modifying adverbs); prepositional phrases; noun phrases (either subject or direct object, with their modifiers); or any set of words beginning with a conjunction, subordinator, or coordinator that expressed more than a simple additive or linking relationship (e.g., a causal link). Two raters independently segmented approximately 15 percent (a little more

than a paragraph) of each of the four texts to assess reliability. There was 92.8 percent agreement between raters in the definition of segments.

The structure of the text was also analyzed to identify the importance or centrality of the various parts of the text. This analysis was based on the principles of top-down text analysis developed by Meyer (1975, 1982, 1985), and her description of *paratactic* (parallel) and *hypotactic* (subordinate) text relations. The unit of analysis was the clause. After segmenting each passage into clauses, we first located all top-level clauses in the text. We began by identifying the thesis statement for the passage, elaborations upon the thesis, and broad organizational concepts that subsumed other information in the passage—that is, any clause that revealed the overall plan or structure of the text (Meyer, 1982). Such clauses often included generalizations about historical events (e.g., that Ike's efforts did nothing to ease Cold War tensions). Meyer (1982) suggests that topic sentences also tend to include this kind of information. We found that topic sentences were a useful starting place for locating top-level information, but due to differences in writing styles, not all topic sentences contained statements that directly elaborated upon the thesis of the passage or provided broad organizational information. Rather, many of the topic sentences focused on specific ideas at a lower level in the structure (which were then elaborated with even lower-level details). All of the clauses that were classified as including organizational information were labeled as A-level clauses.

We next identified the clauses that were subordinate to the A-level clauses. These were clauses that provided details, examples, or specific elaborations of A-level information. These clauses tended to include the "meat" of the passage—the specific details about *who, what, when, where,* and *why.* These clauses conveying mid-level information were classified as B-level clauses.

Finally, the remaining, unclassified clauses were reviewed to ensure that they were all hypotactic to the B-level information. These bottom- or C-level clauses typically included further elaborative information about specific actors and places identified in the B-level clauses (e.g., that Quemoy and Matsu are "a few miles off the coast of China") or colorful anecdotes—what the Time-Life journalists referred to as *nuggets* (Graves et al., 1985). After all clauses had been classified, the structural level of each clause was assigned to all propositions making up that clause.

The analysis of the passages was carried out by a team of three graduate students in rhetoric. Because the classification of clauses required consideration of the specific semantics of the passage, it was not possible to have the classifiers train on one passage and then independently score the experimental passages. Instead, the classifiers as a team worked out the application of the rhetorical principles to the specific text. The classifiers achieved consensus on all classifications; fewer than 10 percent of the classifications required extended discussion to resolve disagreement. The clauses were also reclassified 4 months later in order to assess the reliability of the classifications. That is, we asked two of the classifiers to reclassify a random sample of 13 percent of the clauses in each of the revised texts as an index of test-retest (or, in this case, classification-reclassification) reliability. There was 92.2 percent agreement between the original classification and reclassification 4 months later.

Table 2 presents a summary of the number and proportion of propositions at each structural level in each of the passages. As can be seen in the table, the focus in the original passage was on providing the *who, what, when, where,* and *why,* or middle-level information, with only a little structuring and very little elaboration. The findings of Reder and Anderson (1980) suggest that such an emphasis on summary information with minimal elaboration or structuring would be most effective for facilitating recall. That is, they found that key information is best remembered if it is presented in the form of an unelaborated summary, rather than in a fully elaborated text. However, although two-thirds of the propositions in the original passage were at the middle level, only a third or so of the propositions in the three revisions were devoted

Table 2 Number (and proportion) of propositions at each level of the text hierarchy for four text versions

Text level	Text version			
	Original	Text linguists	Composition teachers	Time-Life
A (Top)	44 (.27)	27 (.15)	75 (.43)	60 (.21)
B (Middle)	109 (.66)	69 (.38)	60 (.35)	92 (.32)
C (Bottom)	11 (.07)	85 (.47)	39 (.22)	137 (.47)
All	164	181	174	289

to conveying this kind of information—the meat of the text.

The composition teachers devoted a considerably greater proportion of the text to providing top-level topic and structure information than did the authors of the original passage (an increase of 60% in the proportion of A-level propositions), whereas the Time-Life revisers and the linguists significantly decreased the proportion of these top-level propositions (by 22% and 44%, respectively). Thus, the revision by the composition teachers was the only one to emphasize the top-level structure of the information presented. Research on advance organizers (Mayer, 1987) and on other means of activating schemata and emphasizing structure (Anderson & Pearson, 1984) suggests that the composition revision should lead to the highest level of learning.

The revisions by the text linguists and the Time-Life writers emphasized elaborations of the middle-level propositions: Almost half of the propositional units in these two revisions were C-level details. Although there is considerable research arguing for the importance of such elaborations (e.g., Bransford, 1979; Mayer, 1979,1984), the work of Reder (1985) suggests that the effectiveness of detailed elaborations depends on exactly how they are used. Any text on the same topic as a given idea may be labeled an elaboration, but it will not necessarily aid in ability to remember or use that idea. For example, if the reader's goal is to remember the main idea, and the main idea is relatively straightforward, then detailed elaborations will distract from the learning task (Reder, 1985). If the reader's goal is simply to remember certain pieces of information, a mnemonic is the most effective learning aid. But if the goal is for the reader to be able to use or apply the main idea, then an effective text will include elaborations that help the reader understand and apply key concepts.

However, the elaborations in the Time-Life version were designed to entertain the reader rather than to help the reader use or apply the main ideas. They tended to stand quite distinct from the core of the text, much like the motivating pictures that are frequently added to text. Therefore, we predicted that these lower-level elaborations would hinder, rather than facilitate, learning of the propositions they elaborated. But the Time-Life revisers argued that the detailed, C-level nuggets they provided would increase interest and provide an anchor for the middle- and top-level propositions (Graves et al., 1985). They argued quite strongly that their elaborations were very important contributors to learning.

Scoring

Short-answer recall data. There were 9 questions testing recall of facts presented in all four passages. We awarded 2 points for each

Table 3 Mean proportion (and number) of propositions recalled on free-recall test by text version and ability

Reading ability	*n*	Text version			
		Original	Text linguists	Composition teachers	Time-Life
High	59	.215 (35.2)	.272 (49.2)	.380 (66.1)	.151 (43.7)
Low	53	.168 (27.5)	.144 (26.0)	.277 (48.1)	.082 (23.6)
All	112	.192 (31.6)	.205 (37.1)	.304 (52.8)	.118 (34.1)

correct answer, and 0 points for each incorrect answer. On the four questions that had two-part answers, 1 point was awarded for each correct part. On the one question that had three parts, we assigned two-thirds of a point for each correct part. (The other four questions were scored either 0 or 2.) Thus, the total score on the short-answer recall test ranged from 0 to 18.

Free recall data. The students' free recalls were segmented into propositions using the same procedure used for analysis of the passages. Scores were compiled for both the number of propositions recalled and the proportion recalled out of all propositions in that passage. We also counted the number and proportion recalled at each level of the text hierarchy. Two raters scored the free recalls; on 12 recalls independently scored by both raters, there was 94% agreement.

Results and discussion

Free recall

Overall performance. Table 3 presents means and standard deviations for the number and the proportion of propositions recalled. We conducted 2 (reading ability) by 4 (text version) between-subjects analyses of variance (ANOVAs) on both the total number of propositions recalled and the proportion of propositions recalled. For the total number recalled, there were statistically significant main effects of both text version, $F(3, 100) = 19.6$, and reading ability, $F(1, 100) = 39.4$, $p < .001$ for both. The interaction effect was not significant, $p > .05$. Tukey HSD post hoc comparisons (Winer, 1971) between the means for the various text versions indicated that significantly more propositions were recalled from the composition teachers' version than from any of the other versions; there was no significant difference between the other versions (critical value $= 8.47$, $p < .05$).

For the proportion of propositions recalled, we again found statistically significant main effects of both text version, $F(3, 100) = 41.9$, and reading ability, $F(1, 100) = 39.1$, $p < .001$ for both. The interaction effect was not significant, $p > .05$. The Tukey comparisons again indicated that a significantly higher proportion of propositions were recalled from the composition teachers' version. In addition, a significantly lower proportion of propositions were recalled from the Time-Life version than for any of the other versions (critical value = .054, $p < .05$). However, the greater length of the Time-Life version may well have accounted for the lower proportion of propositions recalled: There were almost half again as many propositions in the Time-Life version as there were in any of the other texts. Thus, if all of the passages yielded the same absolute number of propositions recalled, the proportion would be much lower for the Time-Life version.

Although the results for proportion recalled may be confounded with passage length, it is nevertheless clear that our results are inconsistent with the Graves et al. finding of superior learning from the Time-Life version. Because the composition teachers' version was shorter than the Time-Life version and about the same length as the other two versions, length alone cannot account for the significantly higher recall found for the composition teachers' version as measured both by the proportion and by the absolute number of propositions recalled.

Analysis of recall by position in text structure. We next examined the number of propositions recalled at each level in the structure of the text. A goal in the design of an instructional text is to facilitate learning of the top- and middle-level propositions. That is, the student should learn the main idea and the supporting detail. The Time-Life revision included highly concrete scenarios at the lowest level of the text structure. We predicted that as a result of this emphasis students would learn the "wrong thing"—that is, they would recall the almost irrelevant details rather than the main ideas. In contrast, we predicted that the composition teachers' version, which emphasized the top-level structure, would lead to better learning of the top- and mid-level propositions.

Figure 1 presents the proportion of propositions recalled out of all propositions at each level of each text. These data were analyzed with a 3 × 4 mixed ANOVA, with text version as the between-subjects variable and level of text structure as the within-subject variable. Reading ability was dropped as a variable in the analysis so as to maximize the number of students contributing to the data (14 free-recall students were missing reading ability data). However, for this and all subsequent analyses, we conducted additional analyses that did include reading ability to determine whether there were any significant interactions with ability. No such interactions were obtained.

We found statistically significant main effects of text version, $F(3, 118) = 19.7$, $p < .001$, and level of text structure, $F(2, 236) = 21.35$, $p < .001$, as well as a significant effect of the interaction between version and level, $F(6, 236) = 2.79$, $p < .05$. Both the composition teachers' version and the Time-Life version contained a large number of propositions at the top level (75 and 60, respectively), almost twice as many propositions as either of the other two versions. Yet the composition teachers' version yielded the highest proportion of top-level propositions recalled, whereas the Time-Life version yielded the lowest proportion. Tukey HSD post hoc comparisons confirmed that a significantly smaller proportion of top-level propositions were recalled from the Time-Life version than from any other version, whereas a significantly higher proportion of top-level propositions were recalled from the composition teachers' version and the text linguists' version than from the other two versions (critical difference = .076, $p < .05$).

The post hoc comparisons also confirmed that a significantly higher proportion of mid-level propositions were recalled from the composition teachers' version than from any other version, and a significantly lower proportion of mid-level propositions were recalled from the Time-Life version than from any other version. Even for propositions at the lowest level in the text structure, a higher proportion were recalled from the composition teachers' version than from either the Time-Life or the original version.

These data suggest that emphasizing the structure of the text does facilitate recall of top-level propositions. Both the composition teachers and the text linguists reported emphasizing cohesion and highlighting the main points in their revision strategy, and in both cases their efforts led to significantly higher recall of those top-level propositions. The Time-Life revisers felt that providing anecdotes would facilitate learning the main ideas. However, their strategy not only did not facilitate recall of those top-level propositions, but in fact resulted in recall of a significantly smaller proportion of mid-level propositions.

However, as shown in Table 2, the four versions all had different total numbers of propositions and different distribution of those

Figure 1
Mean proportion of propositions at each level of the text structure that were recalled

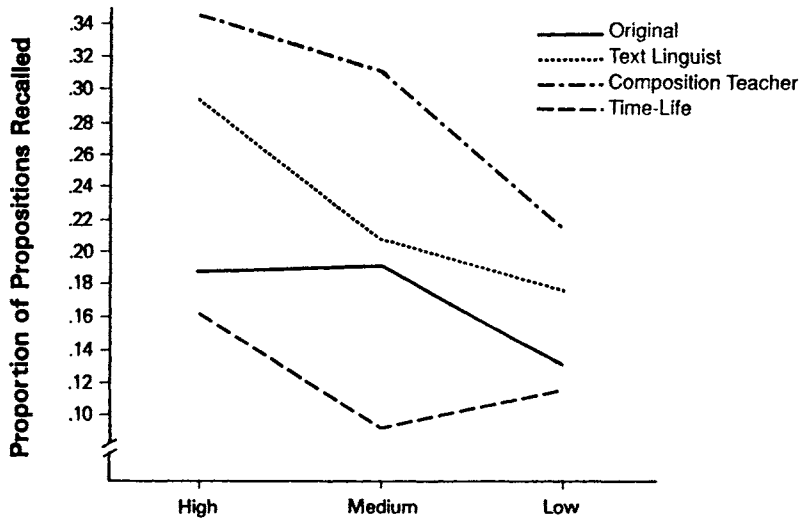
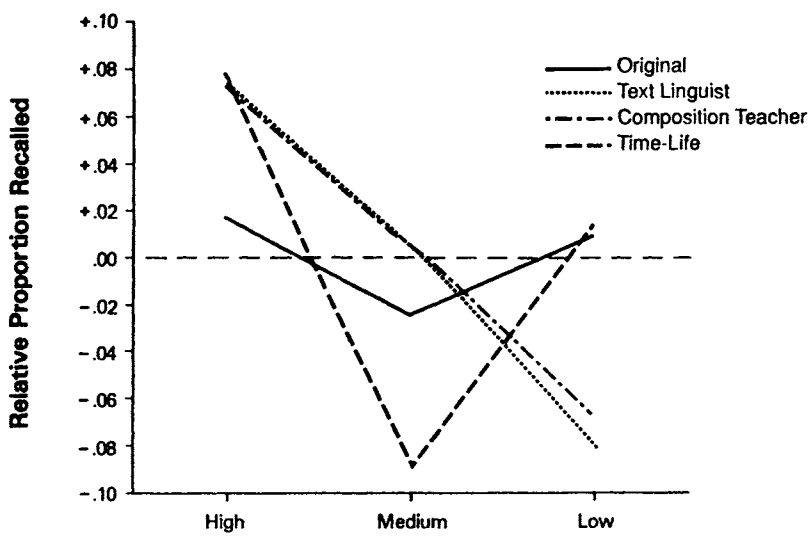


Figure 2
Mean difference between the proportion of total recall that was from each level of the text structure and the proportion of propositions in the text that were at that level

(A negative number means that the students' recalls on average contained proportionally fewer propositions from that level than the proportion in the text.)



propositions across the levels of text structure. Therefore, the results for the proportion of propositions recalled could be confounded with the number of propositions at each level in each text. For example, the high proportion of top-level propositions recalled from the text linguists' version may simply be due to the small number of propositions at that level. Similarly, although students who read the Time-Life version recalled the lowest proportion of C-level propositions, they recalled the highest absolute number of propositions at this level (M = 15.9), because there were more propositions available to be recalled. Therefore, we conducted a second analysis to correct for this possible confound.

Distribution of recall across text structure. The goal of this analysis was still to determine whether certain revisions led to a relatively higher level of recall of propositions that were higher in the text structure. That is, if emphasizing text structure does not make a difference in recall, we would expect the distribution of students' recall across the three levels of text structure to match the distribution of propositions across levels in the text read. If highlighting the structure of the text does affect recall, then in the composition teachers' and text linguists' versions, we would expect the middle- and top-level propositions to comprise a higher proportion of the recall than the proportion they comprised in the text.

We therefore conducted an analysis on the difference between the two distributions of proportions: the proportion of recalled propositions that were at each level of the text structure and the proportion of propositions in the text that were at each level of the text structure. The scores to be analyzed were calculated as follows: The number of propositions recalled at each level of text structure by each student was expressed as a proportion of the total number of propositions recalled by that student. This proportion indexed the relative representation of the different levels of the text structure in the recall profile. However, the proportion of total recall that was at a given level could be large either because the text facilitated learning at that level or because there were more propositions to be recalled at that level. Therefore, we adjusted the proportion scores for each participant based on the distribution of propositions across text levels in the version read (as shown in Table 2). For example, for participants who read the composition teachers' version, the proportion of top-level propositions in the composition teachers' text (.43) was subtracted from the proportion of the propositions recalled that were from the top level.

The mean deviation scores are displayed in Figure 2. The data were analyzed in a 3 (text version) by 4 (level of text structure) mixed (between- and within-subjects) ANOVA. The between-subjects effect in this analysis is zero because the sum of the scores across levels (the sum of a proportional distribution) must equal the whole for each student, and so the difference between subjects will be zero. However, the within-subject effects are unaffected. The analysis yielded a significant main effect of level of text structure, $F(2, 236) = 8.6$, $p < .001$. The level by text version interaction effect approached significance, $F(2, 236) = 2.04$, $p < .06$.

An examination of the data in Figure 2 suggests that the original text did not facilitate recall of the top-level structure, as the deviation scores across all levels of text are very small. In contrast, all three revisers appear to have successfully highlighted the top-level structure of the text. That is, all three revisions resulted in a higher proportion of top-level propositions recalled than the proportion of top-level propositions in the passage. For the text linguists' and composition teachers' versions, the top-level information was recalled at the expense of the information at the lowest level. That is how we would expect learning to occur: remembering more of the main point and forgetting the minor details. However, for the Time-Life version, more top-level information was recalled at the expense of recall of information from the middle level of the text structure. Thus, the vivid anecdotes did not seem to assist the learners in retaining the meat of the passage. Although the proportion of the Time-Life version devoted to low-level propositions was already very high

Table 4 Mean correct score (and standard deviation) on the short-answer test by text version and ability

Reading ability	*n*	Text version			
		Original	Text linguists	Composition teachers	Time-Life
High	55	10.0 (2.1)	10.0 (3.2)	13.9 (3.2)	8.8 (3.1)
Low	55	6.8 (3.5)	5.3 (3.0)	11.0 (4.4)	5.9 (3.1)
All	110	8.5 (3.5)	7.3 (4.0)	12.4 (4.4)	7.5 (3.4)

Note. Maximum score = 18.

(47%), the proportion of low-level propositions in students' recalls was even higher.

Finally, we note that in this analysis the composition teachers' version did not facilitate proportionally more recall of propositions at the highest level than did the other two revisions. Earlier, we had found that 50 percent of the propositions recalled from the composition teachers' version were from the top-level structure of the text, as compared with less than 30 percent for the other versions. Figure 2 suggests that a higher proportion of the recall was at the top level primarily because a much larger proportion of the text was at that level. Thus, the overall effectiveness of the composition teachers' revision was due to a combination of providing a clear text structure and placing a higher proportion of the information at the top level of that structure.

Short-answer recall

Overall performance. Table 4 presents means and standard deviations for scores on the short-answer test. A 2 (reading ability) by 4 (text version) between-subjects ANOVA was performed on these scores. We found statistically significant main effects of both reading ability, $F(1, 102) = 31.3$, and text version, $F(3, 102) = 15.0$, $p < .001$ for both. The interaction effect did not reach significance, $F(3, 102) < 1$. Post hoc comparisons of the four text versions using the Tukey HSD test (Winer, 1971) indicated that performance on the composition teachers' version was higher than on any of the other versions. No other difference exceeded the critical value of 1.23 required for significance at $p < .05$.

Thus, all three measures (short-answer recall score as well as number and proportion scores for free recall) support the effectiveness of the revision by the composition teachers. In contrast, on two measures (short-answer recall and proportion of free recall) the Time-Life revision yielded *lower* performance than any of the other versions. These results are not consistent with the findings of Graves et al., who found that the Time-Life version yielded the highest level of recall, but are consistent with those of Britton et al., who found that the composition teachers' version resulted in the highest recall. The results are also consistent with the majority of research on text comprehension, which indicates that a clear text structure is a critical component of text that promotes effective learning (Mayer, 1987; Meyer, 1984), and that detailed elaborations do not facilitate learning of the main idea unless they aid in the use of that main idea (Reder & Anderson, 1980).

However, the texts were expert revisions, not experimentally generated text, and our main goal in developing questions was to ensure that the questions tapped information present in all

Table 5 Mean proportion correct on short-answer items (and number of items) by location of the answer in the text structure

Location in text structure	Text version			
	Original	Text linguists	Composition teachers	Time-Life
Top	.62 (1)	.70 (1)	.75 (4)	.86 (1)
Middle	.47 (8)	.35 (5)	.61 (4)	.37 (7)
Bottom	– (0)	.32 (3)	.40 (1)	.03 (3)

four texts. Because we used a common set of short-answer questions with texts of quite different structure, there are two possible sources of confound in these data. First, the answers to the questions may have been located earlier in the text for the composition teachers' version than for the other versions, thus capitalizing on a serial position effect in facilitating recall. Alternatively, the answers to the questions might have been located higher in the text structure for the composition teachers' version than for the other versions, making those items easier—because the higher the information in the text structure, the more accessible it should be (Kintsch & van Dijk, 1978; Meyer & Rice, 1984). If our questions chiefly addressed main ideas, then placing the information tapped by the questions higher in the text structure would simply reflect more effective organization on the part of the composition teachers. Nonetheless, we would anticipate that the composition teachers' version would show more effective recall even if we remove the effect of structural level, simply due to their greater attention to cohesiveness. We therefore conducted analyses addressing these two possible confounds.

Analysis of recall by location in passage. To test for differences in the temporal or physical location of the answers, we looked for the relative position of the answers in each text. That is, we first identified the serial location of the sentence or sentences in which each answer occurred. For example, the number of answers in the first five sentences of each version (corresponding to the first paragraph of the original) were as follows: original, 2.6; text linguists, 1.6; composition teachers, 1; and Time-Life, 1. The Time-Life and composition teachers' versions had an equal number of answers in the early portion of the text; therefore, earlier position of answers could not account for the differences in performance between the two versions.

We also calculated the average position of sentences containing all 9 answers for each version. The mean location of answer sentences (average position of the sentence in the text) in the four versions was as follows: original, .51; text linguists, .47; composition teachers, .53; and Time-Life, .60. Although answers tended to occur earlier in the composition teachers' version than in the Time-Life version, the difference is rather small, and the questions for both occurred, on average, later in the text than in either the original or the text linguists' version. Thus, based on the overall evidence, we conclude that the physical location of the answers cannot account for differences in the effectiveness of the versions.

Analysis of recall by position in text structure. We next examined the location of the an-

Table 6 Means (and standard deviations) for student ratings of ease of understanding and enjoyment for four text versions

Rating	Text version			
	Original	Text linguists	Composition teachers	Time-Life
Enjoyment	4.03 (0.91)	4.00 (0.86)	3.12 (1.10)	3.84 (1.14)
Ease of understanding	3.64 (1.00)	3.76 (1.08)	2.78 (1.02)	3.95 (1.08)

Note. Ratings were on a 5-point scale, with 1.0 the highest rating.

swers in relation to the text structure analysis of each version. The number of questions and the proportion correct at each level of the text hierarchy are presented in Table 5.

The data do indicate that recall was directly related to the position of the answers in the text hierarchy; this relationship is consistent across text versions. However, if we look just at the middle level of the text structure, the only level where there are at least two items from each version, we can see that the composition version still yielded a higher level of recall. Thus, structural position of the items alone cannot account for the effectiveness of the composition version.

Interest ratings

The data thus far consistently support the effectiveness of the composition teachers' version of the text in promoting learning and recall. However, students will learn from the text only if they find it sufficiently interesting to keep reading. If they are bored and put the text down, they simply will not learn. This was a major emphasis of the Time-Life revisers; they found the original text dry and boring and could not imagine how the students could be expected to persist with such a text. Therefore, at the end of the experiment, we asked the students to rate how easy it was to learn the information in the text and how enjoyable the text was, relative to the history text they were currently using (or had last used). Eighteen subjects failed to complete the "ease of understanding" item, and 21 subjects failed to complete the "enjoyment" question, leaving sample sizes of 250 and 247, respectively, for these data.

The two sets of ratings were analyzed with 4×2, between-subjects ANOVAs, with text version (the four versions of the text) and type of test (free-recall or short-answer) as the independent variables. (As in previous analyses, reading ability was dropped as a variable to maximize the number of subjects.) Both analyses showed a significant main effect of text version: $F(3, 242) = 15.3$ for ease of learning, and $F(3, 239) = 11.3$ for enjoyment, $p < .001$ for both. Neither the main effect of type of test nor the Test $\times$ Text Version interaction effect approached significance in either analysis. Thus the type of test taken did not influence students' judgment of ease of learning or enjoyment. Examination of the means in Table 6 suggests that the effect of text version in both analyses was due to a higher rating by students who read the composition teachers' version for both enjoyment and ease of understanding (for both questions, the largest difference was between the rating for the composition teachers' version and the ratings for all other versions). The results do not provide any support for the claims of the Time-Life revisers that the original text was boring and that their approach of providing nuggets significantly increased the interest of the

text (Graves et al., 1985). In fact, the Time-Life version was rated only slightly higher than the original text on enjoyment, and lowest of all four versions on ease of understanding.

Conclusion

There is little question that our school textbooks are terribly inadequate. For the most part, the information is uninteresting, poorly organized, and overly broad in coverage (Elliot & Woodward, 1988; Okamoto et al., 1987). How do we provide texts that are more readable and easier to learn from? A recent review of textbooks recommended that publishers should hire writers who can create more vivid stories, rather than hiring social science professionals as writers (Sewall, 1987). Graves et al. make a similar recommendation, based on their data, and news editors who reviewed the Graves et al. text expressed a similar sentiment (Graves et al., 1988). Thus, both popular opinion and the Graves et al. data suggest that vividness and liveliness should be the most important considerations in textbook design.

However, our results suggest that such a conclusion may not be warranted. Our findings, as well as those of Britton et al. (1989), suggest that the version prepared by the composition teachers—not the Time-Life version—led to higher levels of free recall and better performance on a short-answer test. The results reported by Britton et al. further suggest that these effects may have been due to differences in facilitation of retrieval rather than differences in learning. Thus, our findings contradict the popular wisdom that students will find the text enjoyable only if it is vivid and dramatic. In fact, the students in our study gave the text revised by the composition teachers the highest rating for enjoyment as well as for ease of learning. The Time-Life version, on the other hand, was rated only slightly higher than the original text on enjoyment, and lower than the original text on ease of understanding.

Although we can say that the strategy of the Time-Life team did not work to promote learning from the text, the task of saying what *does* work is far more difficult. What are we to tell textbook writers and publishers about the best rhetorical strategy? The experts responsible for the most effective revision strategy—the composition teachers—reported that they emphasized coherence and structure. However, the text linguists also reported that they focused on coherence and structure, yet their text was not as effective in facilitating performance. So we cannot rely only on reported strategies, at least at the superficial level of a retrospective report. Indeed, evidence suggests that there are often significant discrepancies between reported revision strategies and what was actually done (Ericsson & Simon, 1986).

Britton (1986) advocates identifying distinguishing characteristics of the effective or "winning" text. These critical features can then be made available to textbook writers as guidelines for effective writing. Britton et al. (1989) examined the composition teachers' passage and observed that it had the best readability score and also the lowest adjective/verb ratio, which together suggest a less ornate and more active style of writing. They conclude that these may have been the features that resulted in the higher level of comprehension. Our analysis suggests that the composition teachers placed the most emphasis on top-level structure and the least emphasis on the low-level detail. Thus, one might conclude that an extensive top-level structure is essential for learning. However, such a conclusion would be based only on correlational evidence, and thus would be tentative at best. Additional analysis of the texts would probably reveal other differences that are potentially relevant as causal factors.

We are not convinced, however, that isolating relevant variables is the most fruitful approach to developing a strategy to guide textbook writers. For years, authors and educators have presented aspiring writers with guidelines based on isolated characteristics of good writing, but there is little evidence for the effectiveness of this approach. Indeed, if guidelines were effective, we would anticipate that the vast array of guidebooks would have solved the writ-

ing problems in our schools and in our businesses.

In fact, research has provided specific evidence as to the lack of efficacy of guidelines. For example, many states have imposed a readability formula as a guideline for any text to be considered for adoption. Although readability is positively related to learning in naturally occurring text, it seems that using a readability formula as a guideline when writing may lead to less readable school books (Davison & Kantor, 1982). Furthermore, texts revised following readable writing guidelines are seldom easier to understand and may even be more difficult to learn from (Duffy & Kabance, 1982; Pearson, 1974-1975). Outside the educational system, the Department of Defense has formalized the guideline process further by imposing design requirements and guidelines as a contractual requirement in the writing of technical manuals. However, this policy has resulted in complaints from writers that the contractual requirements are frequently inappropriate and are difficult to interpret in the context of specific text (Duffy, Post, & Smith, 1987).

One of the many problems with guidelines is that they are presented independently, as if effectively designed text is simply the summation of a set of individual characteristics. However, designing a comprehensible text is not simply the linear application of guidelines; rather, it might be characterized as an ill-structured problem (Newell & Simon, 1972). That is, there is no one "right" solution to preparing a comprehensible text. The writer does not clearly know when a solution has been achieved. We cannot tell the writer the critical features that a text must possess if it is to be classified as comprehensible, because no one feature is critical. Several writers could effectively improve a given text, each reflecting a different (or partially overlapping) set of guidelines. Similarly, if we gathered a set of well-written texts, we could probably point to a set of features in each text that appear to be related to its comprehensibility. However, it is likely that no one variable would appear consistently across texts or would be effective in improving text comprehension regardless of other text characteristics. It is much like designing a *game* (Wittgenstein, 1953). There are no critical attributes common to all games; thus, we cannot say what characteristics something must have to be classified as a game. One comes to understand the concept of *gameness* through a series of diverse examples, but there will not be a single attribute common to all of those examples.

Spiro, Vispoel, Schmitz, Samarapungavan, and Boerger (1987) have argued that ill-structured domains are typically taught using strategies appropriate only for well-structured domains. That is, even though the knowledge domain is ill structured, there is an attempt to pretend that it is well structured by providing principles to be applied consistently. Spiro et al. argue that it is because such instruction is inappropriate that transfer of learning in ill-structured domains is so low. That is, the learner looks for but cannot locate the structure provided in training, and has not been provided with the skills necessary to build a structure more appropriate to the task at hand. Thus, presenting guidelines for writing that have been abstracted from actual text creates a pretense that writing is a well-structured domain. But when real writing tasks are encountered, the supposedly universal principles—the guidelines—break down.

Spiro et al. argue that, for such ill-structured domains, we must develop cognitive flexibility in the learner. Presenting universal principles, such as guidelines, is counterproductive to the development of flexibility. Cognitive flexibility is best achieved, they argue, by presenting a series of multidimensional examples that reflect the interconnectedness of the features and demonstrate *naturally occurring* complexity and irregularity. If we apply this approach to text design, the texts produced by Graves et al. provide an example and counterexamples of well-formed text. Rather than abstracting the critical features to present to writers, we should present the complete text, highlighting the critical features that distinguish the composition version from the others. Other texts would be used to illustrate other com-

plexes of key features for promoting comprehension. It is by presenting a diverse set of complex, real-world examples and counterexamples that we will best aid the writers of textbooks in producing comprehensible text.

The use of models or complex examples to help in writing and revising text is but one alternative to guidelines. A second alternative that is potentially a very powerful aid to revision is the use of protocol-aided revision. This is a strategy that relies on testing the text with the students for whom it is intended, rather than submitting it to expert judgment as Graves et al. did. In protocol-aided revision, the student is asked to think aloud while studying the text in preparation for a test or for class discussion. The transcript of the student's verbalizations is then used as a basis for revising the text. The value of protocol-aided revision is that it provides detailed data on errors or lapses in comprehension and thus provides more direct guidance for revision (Schriver, 1988a). The use of protocols has been found to be an effective writer-training strategy (Schriver, 1988b). Swaney, Janik, Bond, and Hayes (1981) compared protocol-aided revisions of consumer documents to revisions based on expert judgment. They found little improvement when expert judgment was used alone, but significant improvements when the protocol data were available to guide the revisions.

Protocol-aided revision is but one strategy for conducting *formative evaluation*—testing the materials directly with the audience for whom they are intended. Formative evaluation is becoming common in the computer industry and may have contributed to the general improvement in the quality of those manuals. Unfortunately, it would seem that textbook publishers do not engage in any formative evaluation, much less the detailed evaluation provided by protocol-aided revision. Dick (1987) reports that the state of Florida requires publishers to provide a report on formative evaluation procedures and the resulting text revisions for any text on the state adoption list. The law was ignored during the first 10 years it was in effect, but in 1983 the legislature began to enforce the requirement. However, Dick, in reviewing the reports presented by the publishers, concludes that little formative evaluation is really conducted:

> Relatively little data are collected directly from the learners. Learners are observed during classes, and teachers interpret their reaction to the instruction, but rarely are they tested on the objectives of the instruction or directly interviewed to get their reactions. Clearly this is in stark contrast to both standard formative evaluation procedures and the stated intent of the Florida legislation. (1987, p. 6)

Guidelines, models, and feedback from users all provide product-based information—what the text should or should not look like. It is also important to consider the process—how that information should be used in revising a text. The obvious problem with guidelines is that they provide generalities but do not suggest when or how the principles should be applied. And although protocol-aided revision is itself a method, we still must understand how to make sense of the wealth of information it provides. Research in cognitive science points to the importance of understanding the actual problem-solving behavior of the expert working in his or her domain (Chi, Glaser, & Rees, 1981). Findings from the study of expert writers have had an enormous impact on composition instruction in the schools (Hayes & Flower, 1986). To the best of our knowledge, however, we have as yet no data on the strategies of the expert in the process of revising an instructional text.

Thus, we think that Graves et al. (1988) were on the right track in asking for revisions from experts who reflect different kinds of expertise and testing those revisions with students. With some modifications, procedures such as theirs will contribute to understanding and improving text design. First, it is important to collect data from experts in order to understand their strategies. Unfortunately, the Graves et al. questioning approach yielded more information about the goals of the experts (e.g., to be exciting, to emphasize the main ideas) than about their actual strategies for achieving those goals. Perhaps a fuller understanding of the strategies of such experts would be captured by

data collected while the experts are actually engaged in the process of revising (Ericsson & Simon, 1986). Second, direct testing of students' ability to comprehend the text will contribute to our understanding of effective texts. Again, however, more detailed data on the comprehension process (such as protocols) may yield a fuller understanding of the specific text features that cause difficulties with comprehension. Finally, the various text versions generated by such procedures—along with the many texts gathered and evaluated by Britton and colleagues (Britton, 1986; Glynn & Britton, 1984) and by Graves et al.—will contribute to a library of naturally occurring texts that can be used to illustrate the complex relations between text features. Importantly, however, it is the instantiation of rhetorical principles in naturally occurring text, not the abstracted principles, that will contribute to effective design.

REFERENCES

ANDERSON, R.C., & BIDDLE, W.B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 9, pp. 89-132). New York: Academic Press.

ANDERSON, R.C., & PEARSON, P.D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P.D. Pearson (Ed.), *Handbook of reading research* (pp. 255-291). New York: Longman.

BRANSFORD, J.D. (1979). *Human cognition: Learning, understanding, and remembering.* Belmont, CA: Wadsworth.

BRITTON, B.K. (1986). Capturing art to improve text quality. *Educational Psychologist, 21,* 333-356.

BRITTON, B.K., VAN DUSEN, L., GULGOZ, S., & GLYNN, S.M. (1989). Instructional texts rewritten by five expert teams: Revision and retention improvement. *Journal of Educational Psychology, 81,* 226-239.

BROWN, A.S., & DAY, J.D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22,* 1-14.

BROWN, A.L., & SMILEY, S.S. (1978). The development of strategies for studying text. *Child Development, 49,* 829-835.

CHI, M.T.H., GLASER, R., & REES, E. (1981). *Expertise in problem solving* (Tech. Rep. No. 1981/3). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

DAVISON, A., & KANTOR, R.N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17,* 187-209.

DAVISON, A., & GREEN, G.M. (Eds.). (1988). *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, NJ: Erlbaum.

DICK, W. (1987). *The impact of learner verification and revision on textbook selection in Florida.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

DUFFY, T.M. (1985). Readability formulas: What's the use? In T.M. Duffy & R. Waller (Eds.), *Designing usable texts* (pp. 113-143). Orlando, FL: Academic Press.

DUFFY, T.M., CURRAN, T.S., & SASS, D. (1983). Document design for technical job tasks: An evaluation. *Human Factors, 25,* 143-160.

DUFFY, T.M., & KABANCE, P. (1982). Testing a readable writing approach to text revision. *Journal of Educational Psychology, 74,* 733-748.

DUFFY, T.M., POST, T., & SMITH, G. (1987). Technical manual production: An examination of five systems. *Written Communication, 4,* 370-393.

ELLIOTT, D., & WOODWARD, A. (1988). *Seeing the trees, but missing the forest: Skill and content coverage in elementary textbooks.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

ERICSSON, K.A., & SIMON, H.A. (1986). *Protocol analysis: Verbal reports as data.* Cambridge, MA: Massachusetts Institute of Technology Press.

GLYNN, S.M., & BRITTON, B.K. (1984). Supporting readers' comprehension through effective text design. *Educational Technology, 24,* 40-43.

GRAVES, M.F., & SLATER, W.H. (1986). Could textbooks be better written and would it make a difference? *American Educator, 10,* 36-42.

GRAVES, M.F., SLATER, W.H., ROEN, D.D., REDD-BOYD, T., DUIN, A.H., FURNISS, D.W., & HAZELTINE, P. (1985). *Some characteristics of memorable expository writing: Effects of revisions by writers with different backgrounds.* Symposium presentation at the annual meeting of the American Educational Research Association, Chicago.

GRAVES, M.F., SLATER, W.H., ROEN, D.D., REDD-BOYD, T., DUIN, A.H., FURNISS, D.W., & HAZELTINE, P. (1988). Some characteristics of memorable expository writing: Effects of revisions by writers with different backgrounds. *Research in the Teaching of English, 22,* 242-265.

HAYES, J.R., & FLOWER, L.S. (1986). Writing research and the writer. *American Psychologist, 41,* 1106-1113.

KINCAID, J.P., FISHBURNE, R.P., ROGERS, T., & CHISSOM, B.S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report No. 8-75). Millington, TN: Naval Air Station. (NTIS No. A006 655/5)

KINTSCH, W., & VAN DIJK, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

KNIFFIN, J.D., STEVENSON, C.R., KLARE, G.R., ENTIN, E.B., SLAUGHTER, S.L., & HOOKE, L. (1980). *Operational consequences of a literacy gap* (AFHRL Tech. Rep. No.

79-22). Brooks Air Force Base, TX: Air Force Systems Command.

MAYER, R.E. (1979). Can advance organizers influence meaningful learning? *Review of Educational Psychology, 49,* 371-383.

MAYER, R.E. (1984). Aids to text comprehension. *Educational Psychologist, 19,* 30-42.

MAYER, R.E. (1987). Instructional variables that influence cognitive processes during reading. In B.K. Britton & S.M. Glynn (Eds.), *Executive control processes in reading* (pp. 201-206). Hillsdale, NJ: Erlbaum.

MEYER, B.J.F. (1975). *The organization of prose and its effects on memory.* Amsterdam: North-Holland.

MEYER, B.J.F. (1982). Reading research and the composition teacher: The importance of plans. *College Composition and Communication, 33,* 37-49.

MEYER, B.J.F. (1984). Text dimensions and cognitive processing. In H. Mandl, N.L. Stein, & T. Trabasso (Eds.), *Learning and the comprehension of text* (pp. 3-50). Hillsdale, NJ: Erlbaum.

MEYER, B.J.F. (1985) Prose analysis: Purposes, procedures, and problems. In B.K. Britton & J.B. Black (Eds.), *Understanding expository text* (pp. 11-64). Hillsdale, NJ: Erlbaum.

MEYER, B.J.F., BRANDT, D.M., & BLUTH, G.J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16,* 72-103.

MEYER, B.J.F., & RICE, G.E. (1981). Information recalled from prose by young, middle, and old adult readers. *Experimental Aging Research, 7,* 253-268.

NEWELL, A., & SIMON, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

OKAMOTO, Y., CALFEE, R.C., VARGHESE, S.M., & CHAMBLISS, M.B. (1987, April). *A cross-cultural comparison of textbook designs.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

PAIVIO, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

PEARSON, P.D. (1974-1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. *Reading Research Quarterly, 10,* 155-192.

REDER, L.M. (1985). Techniques available to author, teacher, and reader to improve the retention of main ideas in a chapter. In S.F. Chipman, J.W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (pp. 37-64). Hillsdale, NJ: Erlbaum.

REDER, L.M., & ANDERSON, J.R. (1980). A comparison of texts and their summaries: Memorial consequences. *Journal of Verbal Learning and Verbal Behavior, 19,* 121-134.

RICKARDS, J.P. (1979). Adjunct post questions in text: A critical review of methods and processes. *Review of Educational Research, 49,* 181-196.

SCHRIVER, K.A. (1988a). Document testing procedures: The continuum from text to reader focussed methods. In T.M. Duffy, W. Winn, V. Redish, & K.A. Schriver (Eds.), *Creating usable manuals and forms: A document design symposium* (Tech. Rep. No. 42, pp. 48-70). Pittsburgh, PA: Carnegie Mellon University, Communications Design Center.

SCHRIVER, K.A. (1988b). *Teaching writers to anticipate the reader's needs: A classroom evaluated pedagogy.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

SEWALL, G.T.. (1987). *American history textbooks: An assessment of quality.* New York: Columbia University, Educational Excellence Network.

SPIRO, R.J., VISPOEL, W.P., SCHMITZ, J.G., SAMARAPUNGAVAN, A., & BOERGER, A.E. (1987). Knowledge acquisition and application: Cognitive flexibility and transfer in complex content domains. In B.K. Britton & S.M. Glynn (Eds.), *Executive control processes in reading* (pp. 177-199). Hillsdale, NJ: Erlbaum.

SWANEY, J.H., JANIK, C.A., BOND, S., & HAYES, J.R. (1981). *Editing for comprehension: Improving the process through reading protocols* (Tech. Rep. No. 14). Pittsbugh, PA: Carnegie Mellon University, Communications Design Center.

VAN DUSEN, L., BRITTON, B.K., & GLYNN, S.M. (1987, March). *Retention of instructional texts by three kinds of experts: Test of a retrieval hypothesis.* Paper presented at the Conference on Inference Generation During Discourse Comprehension, Memphis, TN.

WAGNER, R.K., & STERNBERG, R.J. (1987). Executive control in reading comprehension. In B.K. Britton & S.M. Glynn (Eds.), *Executive control processes in reading* (pp. 1-22). Hillsdale, NJ: Erlbaum.

WINER, B.J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

WITTGENSTEIN, L. (1953). *Philosophical investigations.* New York: Macmillan.

Footnote

We wish to thank Mike Benedict for all of his efforts in providing us access to the classrooms and in ensuring that the assessment progressed smoothly.

Received August 2, 1988
Revision received February 22, 1989
Accepted June 15, 1989

APPENDIX

Sample questions and relevant passages from the four text versions

"Who was the leader of Taiwan?"

Original version. Chiang Kai-shek, encouraged to hope for American support in an invasion of mainland China, moved Nationalist troops to Quemoy and Matsu.... In 1954 Dulles negotiated a mutual security pact with Chiang Kai-shek. By this pact, the United States promised to defend Taiwan.... Taiwan agreed not to launch an invasion....

Text linguists' version. The first proposed solution, liberation of the Communist countries, encouraged Chiang Kai-shek, leader of the people of Taiwan, to hope for American support in an invasion of the Chinese mainland. Consequently, Chiang moved Nationalist troops to Quemoy and Matsu.... While occupying the two islands, Chiang negotiated a mutual security pact with Secretary Dulles in 1954. By this pact, Taiwan agreed....

Composition teachers' version. Chiang Kai-shek, the former Chinese leader, had been forced to flee to Taiwan, a small island off the China coast. The American people pressured Eisenhower to take firm action against the Communists in China. Eisenhower ordered Chiang Kai-shek to prepare for an invasion of mainland China.... The U.S. promised to defend Chiang Kai-shek's people in Taiwan.... [T]he Communists left Chiang Kai-shek and the people of Taiwan alone.

Time-Life version. Further south, in the Straits of Taiwan, the Nationalist government of exiled Chinese president Chiang Kai-shek was desperately hanging on under Communist threats of direct attack.... Such bold words encouraged Chiang Kai-shek to act. Hoping that the campaign rhetoric would translate into U.S. support,...he hustled troops from his Taiwanese refuge to Quemoy and Matsu. However, most of the Nationalist refugees on Taiwan lacked their leader's fervor.... Although Chiang Kai-shek did not go ahead for an invasion, he did win a mutual security treaty from the U.S in 1954. This pact committed the U.S. to defend Taiwan.

"What was significant about the outcome of the U.S. presidential election?" OR **"What characteristic of the U.S. president was a major factor in his being elected?"**

(The first version of this question was asked of students who read the Time-Life, original, and text linguists' versions. Students who read the composition teachers' version were asked the second version of the question because the reference in the text was to Eisenhower as a general rather than to Eisenhower as the first Republican president in 20 years.)

Original version. ...election of 1952. The American people elected the first Republican president in 20 years. But the change in power in Washington brought no let up in the tensions of the cold war.

Text linguists' version. ...American presidential election of 1952. In that election the first Republican president in twenty years came to office. But the change of power in Washington brought no let up in tensions of the cold war....

Composition teachers' version. American soldiers were still dying in Korea as the 1952 election approached and naturally, a major issue in the campaign was winning the Korean war. The American people elected Eisenhower, a former general, because he represented a strong military leader who could win the war.

Time-Life version. ...American voters in 1952 elected the hero of World War II, Dwight D. Eisenhower, the first Republican president in twenty years. But the shift in power in Washington failed to calm Soviet-American relations.

References

The Effects of Experience on the Selection of Suitable Retrieval Cues for Studying Texts

Ann L. Brown; Sandra S. Smiley; Sallie Q. C. Lawton

Child Development, Vol. 49, No. 3. (Sep., 1978), pp. 829-835.

Stable URL:

http://links.jstor.org/sici?sici=0009-3920%28197809%2949%3A3%3C829%3ATEOEOT%3E2.0.CO%3B2-4

On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations

Alice Davison; Robert N. Kantor

Reading Research Quarterly, Vol. 17, No. 2. (1982), pp. 187-209.

Stable URL:

http://links.jstor.org/sici?sici=0034-0553%281982%2917%3A2%3C187%3AOTFORF%3E2.0.CO%3B2-5

Reading Research and the Composition Teacher: The Importance of Plans

Bonnie J. F. Meyer

College Composition and Communication, Vol. 33, No. 1. (Feb., 1982), pp. 37-49.

Stable URL:

http://links.jstor.org/sici?sici=0010-096X%28198202%2933%3A1%3C37%3ARRATCT%3E2.0.CO%3B2-Z

Use of Top-Level Structure in Text: Key for Reading Comprehension of Ninth-Grade Students

Bonnie J. F. Meyer; David M. Brandt; George J. Bluth

Reading Research Quarterly, Vol. 16, No. 1. (1980), pp. 72-103.

Stable URL:

http://links.jstor.org/sici?sici=0034-0553%281980%2916%3A1%3C72%3AUOTSIT%3E2.0.CO%3B2-S

LINKED CITATIONS

The Effects of Grammatical Complexity on Children's Comprehension, Recall, and Conception of Certain Semantic Relations

P. David Pearson

Reading Research Quarterly, Vol. 10, No. 2. (1974 - 1975), pp. 155-192.

Stable URL:

http://links.jstor.org/sici?sici=0034-0553%281974%2F1975%2910%3A2%3C155%3ATEOGCO%3E2.0.CO%3B2-K

Adjunct Postquestions in Text: A Critical Review of Methods and Processes

John P. Rickards

Review of Educational Research, Vol. 49, No. 2. (Spring, 1979), pp. 181-196.

Stable URL:

http://links.jstor.org/sici?sici=0034-6543%28197921%2949%3A2%3C181%3AAPITAC%3E2.0.CO%3B2-V