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

P. DAVID PEARSON, UNIVERSITY OF MINNESOTA

ABSTRACT

This study was designed 1) to provide an assessment of linguistic variables which affect the way in which children process verbal data when they read, and 2) to evaluate certain explicit and implicit claims emerging from research and opinion in the areas of readability analysis and psycholinguistics. A repeated measurement design for high and average achieving third and fourth graders yielded data which indicate that grammatical complexity is often an aid to comprehension and recall rather than a hindrance. The results are discussed in light of the implications they provide for writing children's reading material, further readability analysis, and developing models of performance for language users.

Effets des complexités grammaticales sur la compréhension chez les enfants, leur rappel à la mémoire et leur conception de certaines relations sémantiques

Cette étude a été conçue afin de suppléer une évaluation des variables linguistiques qui influent sur la manière dont les enfants assimilent les éléments verbaux. D'autre part, l'étude voudrait évaluer certaines affirmations issues des recherches et des opinions dans les domaines de l'analyse de la lisibilité des textes, et de la psycholinguistique. Des expériences répétées de mesurage, destinées à des écoliers de haut et de moyen accomplissement des 3ème et 4ème années de l'école élémentaire, ont produit des résultats qui suggèrent que la compléxité grammaticale peut souvent aider plutôt que l'enfreindre la compréhension et le rappel à la mémoire. On discute ces résultats en vue de la portée qu'ils pourraient avoir sur la composition des lectures pour les enfants, sur l'analyse de la lisibilité des textes et sur le développement de modèles de performance dans l'emploi de la langue.

Efectos de la complejidad gramatical en la habilidad de los niños para recordar su comprensión y conceptuar ciertas relaciones semánticas

El presente estudio ha sido diseñado 1) para proporcionar una estimación de las variables lingüísticas que afectan el modo en que los niños procesan los datos verbales en la lectura, y 2) para evaluar ciertas aseveraciones explícitas e implícitas, surgidas de la investigación y opinión en las áreas del análisis de amenidad en la lectura y de la psicolingüística. Un diseño de medición repetida para alumnos de tercer y cuarto grado, de mediano y alto rendimiento, proporcionó información que indica que la complejidad gramatical, más que un obstáculo, frecuentemente es una ayuda para su comprensión y recuerdo. Los resultados se discuten a la luz de las implicancias que ellos proveen para escribir material de lectura para niños, promover analisis de amenidad en la lectura, y desarrollar modelos de actuación para aquellos que utilizan idiomas.

This study consists of 3 separate experiments, 2 of which have several parts. Experiment 1 examined, through the use of wh-type questions, the effects of syntactic complexity on children's comprehension of causal relations and of modifying (adjective-noun) relations. Experiments 2 and 3 were conducted subsequent to Experiment 1 in order to shed some light on some of the ambiguous results emanating from Experiment 1.

Experiment 2 examined children's preferences among several syntactically different ways of expressing a common idea. Children were given a question to answer, followed by a set of statements, each of which contained the answer to the question. They were asked to rank the statements according to the degree of helpfulness each provided in answering the question.

Experiment 3 examined differences between the syntactic form in which statements were read by subjects and the syntactic form in which the same statements were later recalled.

PURPOSE OF THE STUDY

Considering the 3 experiments as a group, this study was designed to provide an assessment of linguistic variables which might conceivably affect the way in which children comprehend verbal data when they read. It was simultaneously designed to investigate certain explicit claims and implicit assumptions emerging from research and theoretical positions in the field of transformational-generative grammar.

THEORETICAL POSITIONS UNDER CONSIDERATION

Three theoretical positions are considered as possible candidates in explaining the data obtained in this study as well as that in related studies.

From: *Reading Research Quarterly*, Volume 10, Number 2, pp. 155–192. Copyright © 1974–1975 by the International Reading Association. Reproduced with permission of John Wiley & Sons, Inc.

The first theoretical position is referred to as the *readability hypothesis* because it emanates from assumptions and conclusions stemming from readability research. Broadly speaking, this hypothesis claims that sentence length and sentence complexity contribute to comprehension difficulty.

The second theoretical position is the *deep structure model*. It arises from psycholinguistic research which has been based upon transformational-generative grammars (e.g., Miller and McKean, 1964; Mehler, 1963). It attempts to establish correlates between operations used in transformational-generative grammar and operations used by people in processing verbal data. In short, it attempts to use a grammatical model as a psychological model. The relevant claim of such a model, as it relates to the 3 experiments in this study, is that as surface structure form (the way we see and hear language) approaches deep structure form (the state in which we consciously or unconsciously process and understand language in the mind), comprehension is facilitated. This facilitation occurs because the listener or reader must undergo fewer operations (transformations) in order to analyze, or break down, the surface structure form into deep structure form.

The third theoretical position is referred to as the *chunk model*. It is called the chunk model because it claims that comprehension consists of synthesizing atomistic propositions into larger conceptual or semantic units rather than analyzing complex units into atomistic propositions. If the surface form of a statement is already highly synthesized, comprehension is facilitated. If, on the other hand, the surface form is broken down somewhat (is closer to its deep structure form), comprehension is impeded.

The chunk model and the deep structure model represent diametrically opposed theoretical positions. That which the chunk model predicts will be difficult, the deep structure model predicts will be simple, and vice-versa. The readability hypothesis represents a theoretical position near, but not identical to, the deep structure model. As will be explicated later, sentence length and grammatical complexity tend to co-vary with transformational complexity. That is, longer sentences and sentences with more subordinate clauses and phrases also tend to have more transformations.

DISCUSSION OF RELATED READABILITY RESEARCH

Classic Readability Procedures

Since the readability hypothesis is one of the 3 theoretical positions tested by the data from this study, it is useful to review the procedures used to construct a readability formula.

The classic mode for constructing a readability formula includes these steps:

1. A series of passages known to be graded with respect to difficulty is selected. The basis for grading the passages is usually the number of correct responses made by students judged to have the ability to read at various grade levels to a variety of multiple choice comprehension items accompanying the passages. However, more recently cloze test procedures have been used (e.g., Bormuth, 1966).
2. All potential factors in the passages which might prove to be predictive of passage difficulty are enumerated. They may be formal factors: the number of words per sentence, the number of subordinate clauses, the number of prepositional phrases. Or they may be conceptual factors: the number of words with concrete referents, the number of "vivid" words, the number of abstract words. Usually formula writers have resorted to formal factors because they can be more reliably and objectively measured. Also, since they appear at the surface level, they do not require expert judgment concerning their occurrence. Gray and Leary (1935), who performed the "classic" study in this mode, began with 82 potential formal factors when they set out to develop a formula. They concluded that 44 of the factors were significantly related to reading difficulty. Bormuth (1966), in a more recent attempt, found over 60 formal (or structural) factors that were useful in predicting comprehension difficulty.
3. A multiple regression analysis is performed to determine which factors are most highly related to the criterion measure and at what point the inclusion of another factor in the regression equation ceases to yield a significant increase in the predictive power of the equation.
4. Mathematical transformations are used to translate the formula into grade level equivalents (e.g., 3.2, 4.5, etc.)

Factors Commonly Found in Readability Formulas

While a variety of factors have appeared in different readability formulas, 3 types of factors consistently appear (Klare, 1963). First, almost all formulas have some measure of word difficulty. These usually turn out to be a direct or, more commonly, an indirect measure of word frequency. Second, about 60 per cent of the available formulas use some measure of sentence length. Third, about 30 per cent use some measure of sentence complexity (e.g., number of prepositional phrases or the number of subordinate clauses).

Uses of Readability Formulas

After the regression equations are built, the formula is ready to use as an instrument to measure the difficulty of existing material or as an aid to use in constructing new material. Flesch (1945, 1946a, 1946b, 1951), for example, has prepared several handbooks and sets of recommendations to guide a writer in constructing new materials. While some of his recommendations relate to conceptual elements (abstractness/concreteness), most are methods for reducing sentence length. For example, he recommends using as few adjectives and adverbs as possible and avoiding prepositions and replacing coordinating and subordinating conjunctions with periods.

Such recommendations reveal a common error in interpreting correlational data by assuming that correlation means causality. The fact that sentence length, sentence complexity, or any other factor correlates with the difficulty people experience in answering questions does not imply that altering those correlates will reduce difficulty. It may be that length and complexity are simply indices of complex semantic content; that is, a long or complex

Table 1. Studies Using Reading Comprehension Criteria

Group	Author(s)	Date	Independent Variable	Findings
1	1. Orndorff	1925	Long vs. short sentences	No difference
	2. Gibbons	1931	Long vs. short sentences	Indeterminate
	3. Holland	1933	Long vs. short sentences	Indeterminate
	4. Hites	1950	Long vs. short sentences	No difference
2	5. Nolte	1937	Low vs. high vocabulary scores	No difference
	6. Kueneman	1931	Low vs. high vocabulary scores	No difference
3	7. Marshall	1957	Low vs. high readability scores	No difference
	8. Brown	1952	Low vs. high readability scores	Low scores yielded better comprehension
	9. Klare, Shuford, and Nichols	1957	Low vs. high readability scores	Low scores yielded better comprehension

sentence is long or complex because it represents a concept or principle that could not be communicated in simpler language.

Experimental Studies Based on Readability Formulas

An interesting question to ask is: If you have a concept that you want to communicate, what syntactic form should you select in order to maximize comprehension? Such a question cannot be answered by using correlational analysis. Its answer demands that semantic content be held constant, while syntactic form is varied, between versions of a passage.

Several studies appear to have been designed to answer this question or similar questions. They fall out into 3 groups distinguishable by their independent variables (see Table 1). All the studies in group 1 used sentence length as the independent variable; those in group 2, word frequency; those in group 3, total readability score (an independent variable which included both sentence length and word frequency). The studies are summarized in Table 1. The dependent variable in each was total score on a comprehension test.

An interesting pattern develops. There were no differences between versions when either sentence length or word frequency was the independent variable. The simultaneous application of both, however, appeared to do what neither could do alone.

Unfortunately, there exists no single study which employed a design that permits one to measure the effects of either factor as well as the unique effects due to their interaction.

It is highly unlikely that the studies in group 1 or 3 were adequate measures of the influence of sentence length. Adequacy depends on the kinds of questions that were asked in the comprehension tests. For example, if one rewrites sentence (1) as sentence (2) (which, incidentally, takes Flesch's recommendations), he is manipulating sentence length as a variable. If he uses question (3) to determine the influence of sentence length, he has provided an adequate test of its effect.

(1) Because the dog barked a lot, the boy kicked the dog.
(2) The dog barked a lot. The boy kicked the dog.[1]
(3) Why did the boy kick the dog?
(4) Who kicked the dog?

If, however, he uses question (4), he has tested a relation whose syntactic form is constant across sentences (1) and (2). Question (4) is not relevant to the causal relation whose form is varied between (1) and (2).

Because so many of the classic readability formulas were constructed so long ago, the investigator was unable to uncover any of the tests that were used to grade the passages. Nevertheless, given the general kinds of criteria historically used to build comprehension tests, it seems reasonable to infer that the tests included a variety of types of comprehension questions (e.g., literal comprehension of facts, word meaning, main idea, inferential reasoning, critical reading). If this is a fair inference, then it follows that at least some of the questions would have tested relations whose form was constant across versions (such as question (4)), while some questions would have been so global in nature (e.g., main idea question) that the alterations in form were irrelevant.

In short, it appears unlikely that any of the correlational or experimental studies in readability has provided a fair test of the variables traditionally assumed to influence comprehension difficulty. The question that must be asked to generate an adequate criterion is, "If I have an *idea* I want to communicate, what's the best *way* to communicate it?"

More Recent Readability Research

The most exhaustive readability study in recent years was conducted by Bormuth (1966). He used correlational and multiple regression analyses to determine the predictive power of over 100 structural variables. His study is of special interest because he analyzed a number of variables not used in the classical research and

[1] It can be argued that no causality occurs here. However, causality is often implied in written text rather than made explicit. Hence, this statement represents a real rather than a hypothetical alternative. (Cf., Flesch, 1945, 1946a, 1946b, 1951).

because he used a new criterion to measure passage difficulty: the subjects' ability to fill in cloze tests over the passage.

In general, the same variables that have traditionally shown high correlations with passage difficulty maintained their status. In addition, he found a number of new variables that exhibited high correlations. Several parts of speech ratios were highly related to passage difficulty (e.g., pronoun/conjunction: r = .81; interjection/pronoun: r = .62; verb/conjunction: r = .73). A new measure of sentence complexity based on Yngve's (1960, 1962) word depth analysis was also significantly related to passage difficulty (r = –.55).

Coleman (1965), working with adults, found that relative clauses written in highly embedded forms like (5) were harder to recall than those written in less highly embedded forms, such as (6).

(5) The rat that the cat killed ate the malt.
(6) The cat killed the rat that ate the malt.

He also found that the nominalizations of active verbs (example (7)) were harder to recall than sentences using the active verbs themselves (example (8)).

(7) The boys' planning of the party was a lot of work.
(8) The boys planned the party, and it was a lot of work.

It is difficult to assess the relevance of Coleman's findings to the present study because of his response measure. He had his subjects write down as much as they could remember from passages written in less versus more highly embedded forms and active verb versus nominalization forms. If one tests comprehension with wh-type questions, Coleman's findings might be reversed. For example, given the question,

(9) Which rat ate the malt?

form (5) might well prove to yield better comprehension than form (6).

Hypotheses Emanating from Readability Research

A number of plausible hypotheses concerning the influence of structural factors on comprehension emerge from the classical and recent readability research. If conceptual equivalence is maintained, then a) one longer sentence should prove more difficult than 2 or more shorter sentences; b) the inclusion of subordinating and coordinating conjunctions should increase comprehension difficulty; c) highly embedded forms should be more difficult than low-embedded forms. It is interesting to note that hypothesis b (and to a lesser degree, hypothesis (c) also tests hypothesis (a). Conjunction almost always increases sentence length; embedding usually does.

Recent Psycholinguistic Research

With the advent of transformational-generative grammars (Chomsky, 1957), many psychologists interested in verbal behavior began to look to this new approach to linguistics as a model for explaining language comprehension and production. Beginning with the work of Miller (1960), there have been a host of studies that have attempted in one way or another to make a direct correspondence between the psychological model of the speaker, hearer, or reader and the units and operations of transformational-generative grammars.

Several studies found that transforming kernel sentences (simple active voice declarative statements) into passive, or negative, or interrogative form increases the difficulty that subjects experience in processing the sentences (Miller, 1962; Miller and McKean, 1964; Mehler, 1963). There was a fair relationship obtained between the amount of time taken to process a sentence and the number of transformations involved in getting from the kernel form to the other forms. Gough (1965) tested subjects' ability to verify statements made about pictures placed in front of them. He found that as the number of transformations for a form increased, subjects took longer to verify the statement.

Savin and Perchonok (1965) investigated this issue by using a short term memory task. A sentence was presented along with a string of unrelated words. The subject was instructed to remember the sentence as well as the additional words. Fewer unrelated words were recalled in the case of passive, negative, or interrogative forms than in the case of simple active forms, even though some of the transformed versions were, in fact, shorter than the active versions. In a follow-up study Savin (1966) found that subjects were able to recall more words following right branching forms (10) than following self-embedding forms (11).

(10) The contractor built a house that had three bedrooms.
(11) The house that the contractor built had three bedrooms.

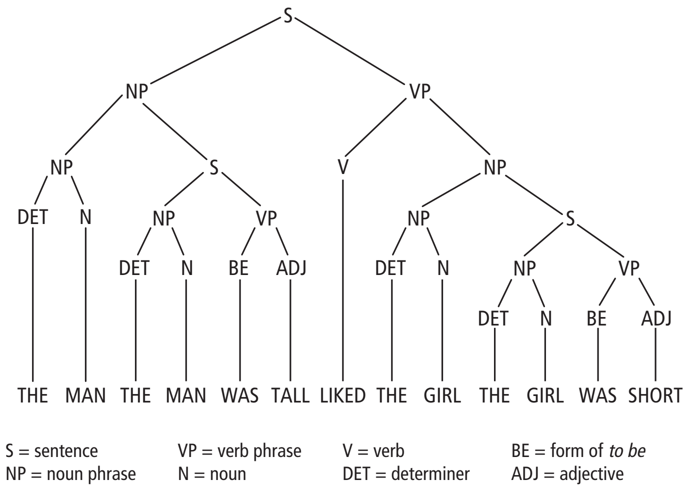
He concluded that the forms involving more transformations or more complex transformations interfered more with memory because they required additional psychological processing in order to get to a deep structure representation.

The relevant point about these studies is that the investigators have explicitly claimed or implicitly assumed a correspondence between grammatical and psychological models. While there is by no means unanimity of opinion regarding these claims (Fodor and Garrett, 1966), much current research operates under the same assumptions. In addition, researchers concerned with the technology of written instruction (e.g., Bormuth, Manning, Carr, and Pearson, 1970; Coleman, 1964, 1965) have conducted experiments that have assumed such a correspondence. It seems useful, therefore, to describe the grammatical bases upon which the corresponding psychological model is founded.

A Sample Transformational Grammar

One of the major distinctions within a transformational generative grammar is between deep structure and surface structure. The meaning of a sentence is represented by its deep structure; the form in which a sentence emerges as speech or writing is represented by its surface structure. The distinction is not trivial because a single surface structure may have more than one possible deep structure, in which case the sentence is ambiguous. Consider the classic example:

Figure 1. Deep Structure Represention of a Sentence with 2 Embedded Adjectives



(12) Flying planes can be dangerous.

It can mean either that you had better stay out of airplanes or that you had better hide when you hear a plane flying overhead. Surface structure ambiguities are resolved at the deep structure level.

A single deep structure may have several possible surface structure representations. Consider the deep structure given in Figure 1. Such a deep structure has several possible surface structure forms. By simply disembedding the 2 sentences under noun phrases, for example, we could get

(13) The man liked the girl. The man was tall. The girl was short.

By embedding *short* as an adjective, we get

(14) The man liked the short girl. The man was tall.

By embedding *the man was tall* as a relative clause, we get

(15) The man who was tall liked the short girl.

By embedding *tall* as an adjective, we get

(16) The tall man liked the short girl.

All of these as well as other possible surface structures differ only in form, not meaning. The semantic interpretation has been determined by the deep structure. Whether or not these surface forms communicate the meaning of the deep structure with equal efficiency is an empirical question. The point is that they are, according to a grammatical interpretation, semantically (or conceptually) equivalent.

In a grammatical analysis, one gets from deep structure to surface structure by performing certain transformations on deep structure. The operations by which we arrived at (13), (14), (15), and (16) are crude statements of transformations.

If one believes that deep structure corresponds in some fashion to the state in which verbal data are processed in the mind, then it is plausible to argue that the number of transformations necessary to get from deep structure to a particular surface structure is an index of the complexity of the surface structure. Hence surface structures exhibiting relatively fewer transformations should be processed with greater speed and accuracy, and vice-versa.

In the set examples (13)–(16), (13) requires the fewest transformations, (14) the second fewest, (15) the third fewest, and (16) the most. If one believes that the grammatical model is a one-to-one mapping of a psychological model (that is, that it describes how data are processed mentally), then it is reasonable to argue that if one tested the efficiency of these forms, they would rank (13) > (14) > (15) > (16). These are exactly the kinds of claims made by the deep structure model.

Recent Research in Conceptual Abstraction (the chunk model)

Within the past 5 years, there has been an attempt by some psycholinguists to offer an alternative view of verbal processing (e.g., Bransford and Franks, 1971). This view hypothesizes that conceptual representational structures or "semantic chunks," rather than atomistic deep structure components, constitute the verbal data which are processed in the mind. Hypothetical storage units might be "tall man," "short girl," or "the tall man hit the short girl." The hearer or reader has to go through some sort of synthesizing process to cement them together (or else he fails to do so and never comprehends the relation.)

Bransford and Franks (1971) have completed several studies supporting the chunk model. They presented adult subjects with larger and smaller components of sentences like (17).

(17) The rock which rolled down the mountain crushed the tiny hut on the edge of the forest.

Later they asked the subjects to state whether or not they had actually heard certain components, and to rate the confidence they had in their judgments. Larger components were given higher recognition scores and confidence ratings than smaller components, irrespective of whether or not they had actually been heard. In other words, the subjects felt more confident about having heard (18) than (19), even though they might have actually heard both of them or neither of them.

(18) The rock which rolled down the mountain crushed the hut.
(19) The hut was tiny.

Bransford and Franks concluded that the findings supported a psychological model that gives primacy to semantic rather than syntactic relations.

In the set of examples (13)–(16), the chunk model would predict that the ranking for comprehension efficiency of the forms would be the exact reverse of the ranking predicted by the deep structure model. Whereas the deep structure would predict (13) > (14) > (15) > (16), the chunk model would predict (16) > (15) > (14) > (13).

REVIEW OF THEORETICAL POSITIONS AND PURPOSES OF THE STUDY

It is clear that the deep structure model and the chunk model stand in opposition to one another. What the one predicts will be simple, the other predicts will be difficult. The third theoretical position, the readability hypothesis, co-varies to a great extent with the deep structure model. For example, when embedded elements are removed from one sentence and expressed as a separate sentence, average sentence length, and, many times, grammatical complexity are reduced.

It should be pointed out that this study has a practical purpose which is relatively independent of its theoretical purpose. That is, it may be possible, using a methodology which holds constant the semantic—or conceptual-nature of a statement while it allows syntactic form to vary, to determine the relative efficiency of various syntactic forms in communicating a given idea. Such a scaling of communication efficiency could prove useful to persons who prepare materials for children, quite independently of whether or not the scaling supported any particular theoretical position.

THE EXPERIMENTS

The data from the 3 separate experiments are reported in this section. Experiment 1 examined children's comprehension of causal and adjectival relations. Experiments 2 and 3 were conducted subsequent to experiment 1 for the purpose of clarifying ambiguity found in the first experiment. Experiment 2 examined children's preference for various syntactical representations of a common idea. Experiment 3 examined children's recall of causal relations.

The situation is further complicated by the fact that experiments 1 and 2 each had 3 parts. In the sections of this report for experiments 1 and 2, the methodology is described for all of experiment 1 and all of experiment 2; however, the results for each of the 3 parts of experiment 1 are reported and discussed separately.

EXPERIMENT 1: COMPREHENSION

Subjects

The subjects were 64 third and fourth grade students attending elementary school in a middle class suburb of Minneapolis. The subjects were selected by participating teachers who were instructed to choose the 5 ablest students from each of their high and middle reading groups. Low ability students were eliminated in order to reduce the likelihood that word recognition problems would complicate measures of comprehension. Initially, 80 subjects were selected. Because of absences, failures to understand the task, and random deletion, the group was reduced from 80 to 64 subjects. The resulting sample was subdivided into 4 groups: 16 medium-achieving third graders (3M), 16 high-achieving third graders (3H), 16 medium-achieving fourth graders (4M), and 16 high-achieving fourth graders (4H). Most of the data were analyzed using grade and achievement level as factors distinguishing between subjects.

Materials

The materials were relatively simple sentences or groups of sentences constructed by the experimenter. They were typed on plain white 4" × 6" index cards in heavy black type. One general criterion was used in generating the materials: that they be as similar as possible to "real" written discourse that children encounter in textbooks and trade books. In order to meet this criterion, the experimenter reviewed children's trade books and basal reading texts to make certain that the sentence types he had chosen for the study represented real alternatives in commonly used materials.

Items for the 3 parts of the comprehension experiment were generated using 4 steps.

Step 1. For each part, decide on the surface structure forms that are of interest as well as the type(s) of wh-questions which provide a fair test of the relation whose surface structure is varied across forms (c.f.,)

For experiment 1.1, eight different surface structure forms were generated by crossing all the combinations—2 levels of each of 3 factors: cue, order and sentence.[2] Table 2 lists these factors in Form Code, explains the levels of each factor, and gives an example of each of the surface structure forms generated by crossing all levels of all factors. In addition, it shows the particular wh-question used to test the relation.

Table 2. Structural Variations in Causal Relations

Form Code*	Example of Form
000	Because John was lazy, he slept all day.
001	John was lazy. So he slept all day.
010	John slept all day because he was lazy.
011	John slept all day. This was because he was lazy.
100	John was lazy and he slept all day.
101	John was lazy. He slept all day.
110	John slept all day, and he was lazy.
111	John slept all day. He was lazy.
–wh	Why did John sleep all day?

* The 3 columns of the form code denote the 2 levels of each factor. The left hand column refers to cuing condition. A cue can either be *present* (0) or *absent* (1). The second column denotes level of order; it can be either *cause-effect* (0) or *effect-cause* (1). The last column denotes sentence level. It can be a *one-sentence* construction (0) or a *2-sentence* construction (1).

For experiments 1.2, and 1.3, four surface structure forms were generated by applying successive transformations on the deep structure representation of a sentence containing 2 embedded sentences which dealt with adjectival relations. Table 3 gives examples of the surface structure forms generated by applying these transformations. It also lists the test questions generated by applying wh-transformations to the deep structure. The *which* question was used in experiment 1.2; the *who* question in experiment 1.3.

[2] The experimenter was unable to locate any transformational-generative grammatical analysis of causal relations as specific and detailed as those available for modifying relations. Hence these operations were generated by the experimenter as a quasi-substitute.

Table 3. Sample Forms for Adjectival Experiments (1.2, 1.3, 2.2, and 2.3)

Form Code	Example of Form
1	The tall man liked the short woman.
2	The man who was tall liked the short woman.
3	The man liked the short woman. He was tall.
4	The man liked the woman. He was tall. She was short.
Which	Which man liked the short woman?
Who	Who liked the short woman?

Step 2. Select as many sentence contents (sentences which contain the relation of interest) as there are surface structure forms. For experiment 1.1 it was necessary to select 8 sentences; for 1.2, four sentences; for 1.3, four sentences.

Step 3. Build a sentence by form matrix by applying each operation outlined in Step 1 to each of the sentences selected in Step 2.

Step 4. To build a test item, select a particular sentence × form combination (a cell in the matrix) and the appropriate test question to go with it. (Note that the test question is the same for all surface structure forms of the same sentence.) Notice that 64 test items are generated in the 8 × 8 causal matrix for experiment 1.1, while 16 test items are generated in each of the 4 × 4 matrices for experiments 1.2 and 1.3.

A test for a given subject was built by assigning test items so that he was exposed to each surface structure form and each sentence once and only once. In experiment 1.1, then, each subject received 8 unique sentence × form combinations (8 unique cells in the matrix). This meant that 8 subjects were needed to gather data on each cell in the matrix—that is, for one complete replication of the matrix. Similarly, 4 subjects were needed to complete a replication of the 4 × 4 matrix in either experiment 1.2 or 1.3.

To control for practice effects and "experimental set," 2 precautions were taken. First, for a given matrix, each surface structure form and each sentence were tested equally often in each serial position within a test. Second, the test items from a given matrix were separated from one another by 7 intervening "dummy" items.

Since the data for each of the 3 comprehension experiments were collected in the same testing situation, items from experiment 1.2 and 1.3 could serve as a portion of the "dummy" items for experiment 1.1, and vice-versa.

Testing Procedures

Each subject was individually tested in an unused classroom relatively free of disturbances. The experimenter told the subject that he was interested in how well children answer questions about what they read. The subject's task was to pick up a card from the tray, read it aloud and hand it to the experimenter. The experimenter then asked the subject a question testing the relation of interest. The experimenter recorded the subject's response on a preconstructed answer sheet. In addition, the entire session was tape recorded so that the experimenter could subsequently verify his response classifications. The subject was tested on 6 practice items before he began the test. Any subject who could not understand the task or who seemed unduly anxious was dismissed from the session.

Treatment of the Data

All of the data from the 3 comprehension experiments were analyzed by using between-subject, within-subject analyses of variance for dichotomous data (Winer, 1962). Two between factors, grade level (3 or 4) and achievement level (medium or high), were common to analyses from experiments 1.1, 1.2, and 1.3. The within-subjects factors (operations used to generate surface forms) as well as the scoring procedures for experiment 1.1 (causal relations) differed from those used in experiments 1.2 and 1.3.

For experiment 1.1 the 3 within-subjects factors were cue, order, and sentence level (c.f., Table 2). Two dependent variables were used: number of correct responses and number of subordinate responses. A response was scored correct if it contained the major lexical elements in the dependent clause of form 000. Reasonable paraphrases were also scored correct. Using the example in Table 2, (20), (21), and (22) would have been scored correct.

(20) because John (or he) was lazy.
(21) John (or he) was lazy.
(22) Lazy people sleep all day.

A response was scored as subordinate if it began with the word *because* or a reasonable semantic substitute for *because* (since, as, for). Subordinated responses were examined in order to assess the stability of response outputs as a function of the varying stimulus inputs.

For the experiments 1.2 and 1.3, surface structure form (the 4 forms shown in Table 3) served as the within-subjects independent variable. In each case, form 1 was always the form representing the greatest number of transformations (from deep to surface structure), while form 4 represented the fewest. The overall statistical test for *form* was omitted, allowing these orthogonal contrasts:

$$\Gamma_1 = F_1 + F_2 - F_3 - F_4$$
$$\Gamma_2 = F_1 - F_2$$
$$\Gamma_3 = F_3 - F_4$$

For experiment 1.2, the adjective experiment which used a *which* question to test the effect of structural changes, there were 3 dependent variables: a) number of *errors*, b) number of *adjectival* responses, and c) number of *clausal* responses. The 3 dependent variables constituted mutually exclusive categories. A response was scored as *adjectival* if it was of the form, "the *tall* boy," or "the *tall* one." A response was scored as *clausal* if it was of the form, "the boy *who was tall*," or "the one *who was tall*." The following kinds of responses were scored as *errors*: a) incorrect adjectives or clauses assigned to the nominal, b) no response, c) any otherwise unclassified response.

For experiment 1.3—the adjective experiment which used a *who* question—the dependent variables were slightly different from those used in experiment 1.2. They included number of *errors*, number of *adjective-noun* responses (like the *adjective* responses in 1.2), number of *noun-clause* responses (akin to the *clausal* responses in 1.2), and number of *noun* responses. The classification procedures were identical to the adjective-*which* study, except for *noun* responses. A response was classified as a *noun* if the correct *noun* was used without either an adjectival or a clausal modifier.

Results and Discussion

Experiment 1.1—causal relations. When the number of correct responses is analyzed, the surprising finding is that differences between groups, among forms, or among cells are so small. The largest difference between groups is 2; the largest between forms, also 2. In short, virtually every subject responded correctly to every form. Out of a total of 512 responses, there were only 11 errors. No analysis of variance was computed.

A different picture results when the dependent variable is number of subordinated responses. Cell totals are reported in Table 4. There was a significantly higher total for the *cue-present* (0++)[3] condition (T = 193) than for *cue-absent* (1++) condition (T = 137), $F_{(1,60)} = 26.3750$, $p < .01$. The difference between *one-sentence* (++0) forms (T = 182) and *2-sentence* (++1) forms (T = 148) was also significant $F_{(1,60)} = 16.5034$, $p < .01$. Differences for other main effects were not significant.

Table 4. Cell Totals: Experiment 1.1—Number of Subordinated Responses

	Form								
Group	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**	**T_{Group}**
3M	11	8	14	13	7	7	10	8	78
3H	13	9	15	14	10	11	9	7	88
4M	11	8	13	13	9	9	10	7	80
4H	16	10	12	13	12	7	10	4	84
T_{Form}	51	35	54	53	38	34	39	26	330

Of the 26 interactions, only 2 were significant: a) cue × order, $F_{(1,60)} = 17.6667$, $p < .01$; and b) cue × order × sentence, $F_{(1,60)} = 8.2640$, $p < .01$. The cue × order interaction graph indicates that when a cue was present, the *effect-cause* order produced more subordinated responses, but that when no cue word was present, the *cause-effect* order produced more. This interaction was really due to the unique influences of forms 001 and 111. Form 001, the *so* form was the only form within the *cue-present* condition that was different from the others. It did not contain the subordinating conjunction *because*. Form 111 likewise depressed the totals for the *effect-cause* order within the *no-cue* condition. An examination of the cue × order × sentence interaction revealed that these same 2 forms, 001 and 111, are mainly responsible for the interaction.

It is clear that, in terms of number of correct responses, no support for the readability hypothesis or the chunk model is possible. The surprising finding is how well, not how poorly, all groups did on all forms. Perhaps the semantic content of the sentences was so simple that it masked possible differences due to form.

In terms of the number of subordinated responses (those that begin with *because* or one of its synonyms), there are clear effects due to conditions of cuing and sentence; the cuing effect is even more striking if one compares form 001, the *so* form, with the other *cue-present* forms. The other 3 all contained the word *because*; and they all elicited a higher number of *because* responses than the *so* form. This difference is not surprising because, given a *why* question about a causal relationship, there is no reason to begin the response with *so.* It is syntactically and logically unnecessary. The more interesting fact is that there were as many subordinated responses as there were to the *so* stimulus condition and to the various *cue-absent* stimulus conditions.

Despite the cue × order × sentence interaction, the sentence effect is in the same direction across cue × order conditions. If one regards each successive pair of forms as minimal pairs differing only with respect to sentence condition, the difference between members within each pair favor the *one-sentence* condition. (See Table 5.) It is true, however, that the differences between members of a pair vary widely between pairs.

Table 5. Differences Due to Sentence Condition Within and Between Pairs Classified by Cue and Order Conditions: Experiment 1.1

		Pair							
		1		2		3		4	
Code	**Cue × Order** **Sentence**	00_ __0	00_ __1	01_ __0	01_ __1	10_ __0	10_ __1	11_ __0	11_ __1
Totals		51	35	54	53	38	34	39	26
Differences		16		1		4		13	
Differences obtained by subtracting (__1) from (__0).									

It is somewhat difficult to evaluate what it is that the dependent variable of number of subordinated responses means. It is clear that (23) is just as correct an answer as (24); perhaps the insertion of *because* indicates a more *unified* conception of the causal relation.

(23) John was lazy.
(24) Because John was lazy.

If that is a reasonable view, it follows that both cuing and sentence conditions have an effect on children's ability to unify a causal relation, but that the effects do vary across levels of order (that is, the cue × order × sentence interaction). At best, however, this is a speculative explanation. However, data from experiments 2.1 and 3 do shed light on this explanation.

[3] Numbers refer back to Table 2. The + signs indicate that we are summing over levels of these variables. The first of the 3 numeral positions indicates cuing variable; the second, order; the third, sentence.

Table 6. Cell Totals: Experiment 1.2—Number of Responses by Form and Group

		Form				
Response Type	Group	1	2	3	4	T_{Group}
Errors	3M	5	1	4	3	13
	3H	2	1	4	4	11
	4M	3	3	6	5	17
	4H	2	0	4	6	12
	T_{Form}	12	5	18	18	53
Adjective Responses	3M	11	7	7	9	34
	3H	14	2	7	5	28
	4M	13	6	8	10	37
	4H	14	2	7	7	30
	T_{Form}	52	17	29	31	129
Clausal Responses	3M	0	8	5	4	17
	3H	0	13	5	7	25
	4M	0	7	2	2	11
	4H	0	14	5	3	22
	T_{Form}	0	42	17	16	75

Experiment 1.2—adjective relations (using a which *question).* The cell totals for the 3 dependent variables are reported in Table 6. The ANOVA for the number of *errors* revealed that none of the between-subject effects and none of the interaction effects were significant. The only significant comparison within subjects was the contrast:

$$\Gamma_1 = F_1 + F_2 - F_3 - F_4$$
$$12 + 5 - 18 - 18$$
$$F_{(1,\ 180)} = 8.2904,\ p < .01.$$

F_1 and F_2, the highly cohesive forms, yielded significantly fewer errors than the less cohesive forms, F_3 and F_4. The comparison between the adjective form (Form 1) and the clausal form (Form 2) favored the clausal, but the difference was not significant.

The ANOVA for the number of *adjectival responses* indicated that none of the between-subject comparisons was significant.

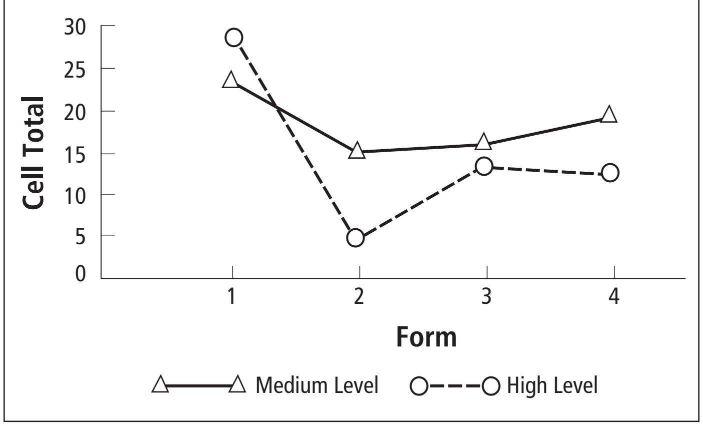
Of the 3 specific orthogonal contrasts between levels of form, only the contrast Γ_2 was significant:

$$\Gamma_2 : F_1 - F_2$$
$$: 52 - 17$$
$$F_{(1,180)} = 46.6113,\ p < .01.$$

There was a high incidence of *adjectival* responses to both F_3 and F_4 stimuli. One interaction was significant: level × form, $F_{(3,\ 180)} = 2.1705$, $.01 < p < .05$. The interaction graph, Figure 2, indicates that high achievers gave more *adjectival* responses to adjective (F_1) stimuli, but that medium achievers gave more *adjectival* responses to other stimulus forms.

The results for the dependent variable, number of *clausal* responses, revealed a significant effect due to level ($T_M = 28, T_H = 47$), $F_{(1,60)} = 6.0368$, $.01 < p < .05$. The contrast between the adjective and clausal stimulus forms was also significant:

Figure 2. Graph of Level X Form Interaction: Experiment 1.2: Number of *Adjectival* Responses



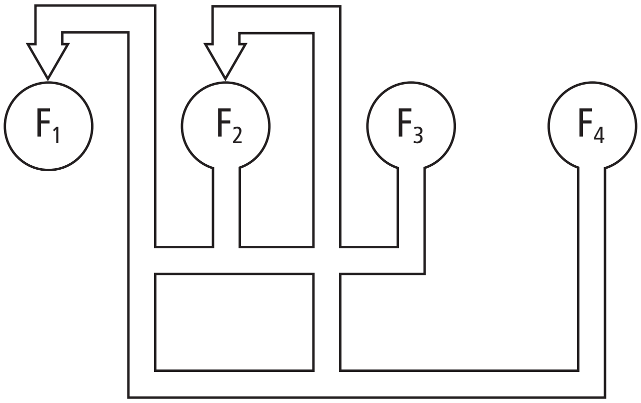
$$\Gamma_2 : F_1 - F_2$$
$$: 0 - 42$$
$$F_{(1,180)} = 69.7667,\ p < .01.$$

Only one of 6 between-subject comparisons was significant. There are few data to indicate any important developmental trends in the response outputs made by the different groups of subjects. There is a tendency, however, for high achievers to give relatively more *clausal* responses, while medium achievers seem to give more *adjectival* responses.

All 3 dependent variables were sensitive to differences between forms, but none of them matched the predictions outlined by the transformational model. The error data lend partial support to the chunk model, but they separate the forms into 2 rather than 4 categories. Comprehension of the *which* question is better when the relationship is stated more cohesively (F_1 or F_2) than when it is stated less cohesively (F_3 or F_4), but there are no significant distinctions within these categories. The lack of a distinction between F_3 and F_4 is reasonable because the relationship tested by the *which* question was constant across these forms. A relative clause structure (F_2) appears, however, to provide as much, if not more, cohesiveness for the nominal-modifier relationship than does an adjective structure (F_1).

The other dependent variables, number of *adjectival responses* and number of *clausal responses*, provide predictable results. Within the 2 highly cohesive forms (F_1 and F_2), there is a strong tendency for responses to match the stimulus input. While there is a small amount of crossover from a clause input to an *adjectival* output (response), there is no crossover in the opposite direction. The crossover from less cohesive inputs (F_3 and F_4) was proportional for both *adjectival* and *clausal* responses. If the 4 forms are placed on a continuum from the theoretically most cohesive to the theoretically least cohesive (Figure 3), then output flow is unidirectional; that is, it goes only from less cohesive inputs to more cohesive outputs. Thus both the error data and the correct response data lend some support to the chunk model. Responses move toward a more cohesive form because the subjects' perception of the relationships involved converges on the more cohesive forms.

Figure 3. Diagram of Response Output Flow from Stimulus Inputs: Experiment 1.2



Experiment 1.3—adjective relations (using a who *question).* The unidirectional flow in Figure 4 presents an interesting, but biased, picture. It is interesting because the flow is unidirectional, indicating that increasingly cohesive forms are more stable in terms of output. But it is biased because there is no crossover possible from F_1 or F_2 or F_3 or F_4; that is, the *which* question preempts a *noun* response.

When a set of materials identical in form to those in experiment 1.2 are constructed, and when the relationship tested is the *nominal-rest of the sentence* relationship, the data provide a more reasonable test of the flow model in Figure 4. This results from the fact that, given a *who* question, a single *noun* ("the *boy*") is equally as reasonable a response as an *adjectival* ("the *tall boy*") response or a *clausal* ("the boy *who was tall*") response. Note, however, that a price is paid for this test: the *who* question does not test the nominal-modifier relationship, according to the criteria established earlier (see Figure 1). Even so, it is a useful study because it allows one to look at the stability of different forms as measured by the crossover tendencies of responses elicited from stimulus inputs.

The results from experiment 1.3 are reported in Table 7. In general, there were no significant effects due to either grade or level; nor were there any significant interactions. The cell totals for number of *errors* are notable for their infrequency rather than their frequency. Of the 256 responses classified, only 5 were classified as *errors*.

The ANOVA for number of *adjective-noun* responses revealed 2 significant contrasts between levels of form.

a] $\Gamma_1 : F_1 + F_2 - F_3 - F_4$
$: 61 + 8 - 16 - 13$
$F_{(1,\,180)} = 50.6483,\ p < .01.$

b] $\Gamma_2 : F_1 - F_2$
$: 61 - 8$
$F_{(1,180)} = 177.8387,\ p < .01.$

Clearly, F_1 elicited far more *adjective-noun* responses than any other form. With respect to the number of *noun-clause* responses, 2 contrasts between levels of form were significant:

Table 7. Cell Totals: Experiment 1.3—Number of Responses by Form and Group

		Form				
Response Type	Group	1	2	3	4	T_{Group}
Errors	3M	0	0	0	3	3
	3H	0	0	0	0	0
	4M	1	0	0	1	2
	4H	0	0	0	0	0
	T_{Form}	1	0	0	4	5
Adjective-Noun Responses	3M	16	0	5	1	22
	3H	15	1	3	4	23
	4M	15	3	3	3	24
	4H	15	4	5	5	29
	T_{Form}	61	8	16	13	98
Noun-Clause Responses	3M	0	13	4	4	21
	3H	0	15	5	3	23
	4M	0	11	4	2	17
	4H	0	12	3	2	17
	T_{Form}	0	51	16	11	78
Noun Responses	3M	0	3	7	8	18
	3H	1	0	8	9	18
	4M	0	2	9	10	21
	4H	1	0	8	9	18
	T_{Form}	2	5	32	36	75

a] $\Gamma_1 : F_1 + F_2 - F_3 - F_4$
$: 0 + 51 - 16 - 11$
$F_{(1,\,180)} = 19.4546,\ p < .01.$

b] $\Gamma_2 : F_1 - F_2$
$: 0 - 51$
$F_{(1,180)} = 175.7811,\ p < .01.$

The results obtained when the dependent variable was number of *noun-clause* responses were nearly the perfect inverse of the results obtained when the dependent variable was number of *adjective-noun* responses.

The ANOVA for the number of *noun* responses indicated that one contrast was significant:

$\Gamma_1 : F_1 + F_2 - F_3 - F_4$
$: 2 + 5 - 32 - 36$
$F_{(1,\,180)} = 99.1480,\ p < .01.$

The least cohesive forms elicited far more *noun* responses than the more cohesive forms. The differences between the 2 less cohesive forms or between the 2 more cohesive forms were small and insignificant.

The error data provide no basis for distinguishing between any factors. There are too few errors on which to make any judgments. The data from the other 3 dependent variables are consonant with the data from experiment 1.2, and the crossover patterns from stimulus input to response output provide support for the model diagrammed in Figure 6. Some modifications are necessary, however, because there is at least some crossover from

Table 8. Input-Output Matrix: Experiment 1.3

		Form of the Response			
		Adj-Noun	Noun Clause	Noun	Error
Form of the Stimulus	F_1 Adj-Noun	61	0	2	1
	F_2 Noun Clause	8	51	5	0
	F_3 $Noun_1$	16	16	32	0
	F_4 $Noun_2$	13	11	36	4
	Total	98	78	75	5

more cohesive inputs to less cohesive outputs. Table 8 provides a useful format for scrutinizing the data. Stimulus inputs are listed in the rows; response outputs are listed in the columns.

First of all, the data do not discriminate very well between F_3 and F_4 inputs. Apparently once the modifier is formally removed from the nominal, no further psychological separation occurs. If F_3 and F_4 are considered as a single state, then it appears that as input forms become less cohesive, stability (the incidence of matches between input and outputs) decreases.

The general trend of response flow convergence is toward more cohesive forms for those responses that do not match stimulus inputs. The crossover toward less cohesive forms is minimal.

Taken as a unit, experiments 1.2 and 1.3 indicate that more cohesive forms yield better comprehension and more stable comprehension. Furthermore, subjects' preception of the stimulus inputs for adjectival relationship move toward the more cohesive response outputs. The chunk model can accommodate these findings more easily than the deep structure model. In fact, they are not predictable at all under the deep structure model. The only question about the chunk model is whether or not there are really 4 levels of cohesiveness. The data would seem to indicate that 2, possibly 3, levels are more reasonable.

EXPERIMENT 2: PREFERENCE STUDIES

Because of the failure of the comprehension experiments to yield an unambiguous interpretation of the models and in order to determine whether or not important effects in the comprehension experiments were generalizable across response modes, a follow-up preference study was conducted.

Subjects for Experiment 2

The subjects were 24 fourth grade students randomly selected from 2 fourth grade classrooms not involved in the comprehension study. All readers in these classrooms were reading materials at grade level. Hence it is unlikely that word recognition problems interfered with reading the comparatively easy test items.

Materials for Experiment 2

The materials were developed in essentially the same manner as in the comprehension study; that is N × N sentence × form matrices were constructed. However, the method for extracting an item was quite different. The subjects' task was to choose which form, from among N forms, he preferred when all forms were represented within an item. Therefore, items were distinguished only by sentence content. In the comprehension study, a 4 × 4 matrix generated 16 items; but in the preference study, a 4 × 4 matrix generated only 4 items.

The factor of order (C-E or E-C) was eliminated from the causal comprehension matrix, yielding a 4 × 4 rather than an 8 × 8 matrix. The experimenter felt that the subjects would not be able to rank 8 forms reliably. Besides the items generated for the adjective studies and the causal study, buffer items were generated by including 5 other 4 × 4 sentence × form matrices. Each subject received a test booklet containing 36 items. The 4 items from a particular matrix were separated from one another by 8 intervening items. The serial position for a given form was rotated between items so that every form appeared in every position.

A sample item for the experiment 2.1 (causal relations) was

Why did John sleep all day?
___1. Because John was lazy, he slept all day.
___2. John was lazy. So he slept all day.
___3. John was lazy, and he slept all day.
___4. John was lazy. He slept all day.

A sample item for experiment 2.2 (adjective relations which) was

Which man thanked the young woman?
___1. The tall man thanked the young woman.
___2. The man who was tall thanked the young woman.
___3. The man thanked the young woman. He was tall.
___4. The man thanked the woman. He was tall. She was young.

A sample item for experiment 2.3 (adjective relations who) was

Who thanked the young woman?
___1. The tall man thanked the young woman.
___2. The man who was tall thanked the young woman.
___3. The man thanked the young woman. He was tall.
___4. The man thanked the woman. He was tall. She was young.

Testing Procedures for Experiment 2

The preference testing was carried out in a group situation. The experimenter told the subjects that their task was to rank the forms in terms of their perceptions about the relative clarity and simplicity of the forms; that is, how much help they thought each form would provide in answering the question listed at the beginning of each item. To the form the subjects considered the best, easiest, and clearest, they assigned a rank of 1; to the second best, easiest, and clearest, a rank of 2; for the third, a rank of 3; for the worst, hardest, and least clear, a rank of 4. The experimenter explained the task to the subjects, conducted 4 examples with the entire group, and circulated among the subjects to make certain that each one understood the task.

Treatment of the Data

Ranks for each form summed across the 4 items within a matrix for each subject. The resulting sum was used to rerank the forms. When ties occurred, average ranks were assigned. For example:

If the summed ranks for a subject were:	Then the assigned ranks for that subject were:
Form 1–8	1
Form 2–10	2.5
Form 3–10	2.5
Form 4–12	4

The assigned ranks were subjected to a Friedman χ^2 analysis of ranks (Winer, 1962) to determine differences due to form.

Results and Discussion of Experiment 2

The sums of ranks and the mean ranks for each of the forms are reported in Table 9. The differences attributable to form are significant for each of the 3 sets of ranks:

Causal—$\chi^2_{(3)} = 8.6266, .025 < p < .05$
Adjective which—$\chi^2_{(3)} = 29.5107, p < .001$
Who—$\chi^2_{(3)} = 34.7125, p < .001$

The magnitude of the test statistics for the adjective data indicate a clear trend on the part of subjects to select the more cohesive, more heavily embedded forms as preferable to the less cohesive, less heavily embedded forms.

The trend is not nearly as impressive for the data in the causal study; however *one sentence* forms are preferred to *two sentence* forms in either cuing conditions, and *cue-present* forms are preferred to *cue-absent* forms in either sentence condition.

A revealing picture results when the frequency of ranks between forms for the causal is graphed. The data in Figure 4 are even more striking than the summed ranks in Table 9. Ranks #1, #2, and #3 each predict the "closeness" of the 4 forms as measured by the sums of ranks.

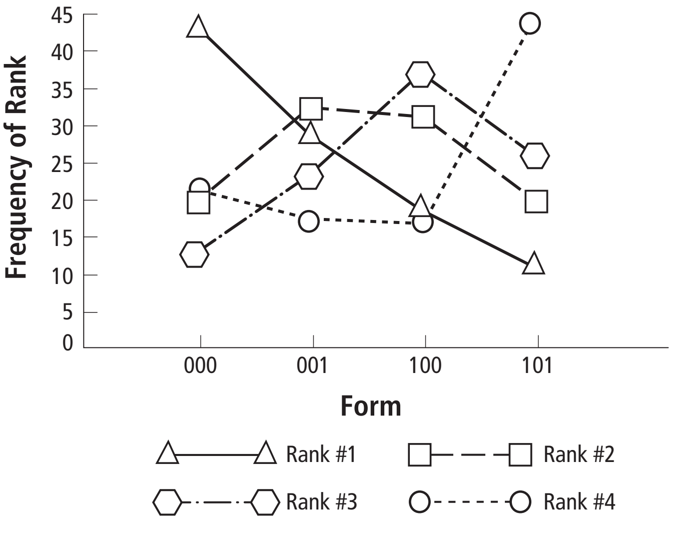
Table 9. Sums of Ranks Between Forms: Experiments 2.1, 2.2, 2.3

Experiment 2.1
Causal

Code	Form	Σ ranks	M rank
000	*because*	49.0	2.04
001	*so*	54.5	2.27
100	*and*	64.0	2.67
101	*no cue*	72.5	3.02

	Experiment 2.2		Experiment 2.3	
	Which Study		Who Study	
Form	Σ ranks	M rank	Σ ranks	M rank
1	36.0	1.50	35.5	1.49
2	56.0	2.33	51.5	2.15
3	65.0	2.70	71.5	2.98
4	83.0	3.46	81.5	3.39

Figure 4. Frequencies of Ranks Between Forms for the Causal Relation: Experiment 2.1



Only rank #4 does not discriminate along a scale. It appears to classify the 4 forms into 2 rather than 4 categories. This probably happens because the subjects rank from best (#1) through worst (#4). Hence, they may have run out of discriminating power by the time they get to #4.

Clearly, the data indicate a marked preference for the more cohesive, more grammatically complex forms. These findings can be accommodated quite easily by the chunk model. They lend no support to either the readability hypothesis or the deep structure model.

EXPERIMENT 3: AIDED RECALL OF CAUSAL RELATIONS

In order to shed light on some of the ambiguous results obtained in experiment 1.1, a modest follow-up experiment was conducted. It was felt that an examination of the relationship between stimulus input and recall output might help to explain how subjects deal with relations of causality.

Method for Experiment 3

The subjects were 8 fourth grade students who had not been used in the comprehension study. All were judged by their teachers to be average or high achievers; however, achievement level was not used as a blocking variable.

The materials were exactly the same as those used in the causal-comprehension set. In addition, 16 buffer items were included to separate the causal forms. They were selected from among the other sets in the comprehension study.

Each subject was tested individually. The experimenter told him to read each sentence and to try to remember it as best he could, because later on he would be asked to recall it. After the subject had read all 24 sentences, the experimenter began the aided recall procedure. It proceeded in this fashion: Assume that the recall item was

(25) John was lazy, and he slept all day.

The first cue was always the first lexical item in the sentence, in this case, *John*. The experimenter asked, "Do you remember the sentence about *John*?" If the subject did, the experimenter recorded the form in which he recalled it. If he did not, the experimenter gave the second lexical item as a cue; in this example, *lazy*. The third cue was the third lexical item, *slept*. If the subject did not recall the sentence after 3 cues, the experimenter went on to a new item. The form of the response was recorded by the experimenter.

Results and Discussion of Experiment 3

The results from the aided recall study indicate very strong influences due to cuing condition and sentence condition. Because of the small number of subjects and because the results were so clear cut, the data were not subjected to inferential statistical analysis. Instead they are summarized in Table 10 in the form of an input-output matrix. The input denotes the form in which the subjects read the statement; the output denotes the form in which they recalled it. The notation for forms are the same as those used in experiment 1.1 and shown in Table 2.

A measure of input-output stability is indicated by the diagonal from upper left to lower right. Numbers in the diagonal indicate a direct match between input and output. Only forms 000, 001, 010 were stable. All other inputs, for the most part, were very unstable.

The influence of cuing condition is overwhelming. For *cue-present* inputs there were 29 *cue-present* outputs and no *cue-absent* outputs. Of the 3 errors, 2 resulted from a form 011 input, an extremely unstable form. As a matter of fact, there were no outputs in form 011 regardless of input form.

On the other hand, for *cue-absent* inputs, there were 21 *cue-present* outputs. Furthermore, more errors (T = 6) than *cue-absent* outputs (T = 5) resulted from *cue-absent* inputs.

The effect or order is also quite striking. No *cause-effect* input was recalled as an *effect-cause* output, but 7 *effect-cause* inputs resulted in *cause-effect* outputs. In addition, the *effect-cause* order yielded 7 errors as opposed to only 2 for the *cause-effect* order.

The sentence factor also proved to be striking. There were no differences in number of errors. However, there was stability for *one-sentence* inputs (22 out of 32) than for *2-sentence* inputs (10 out of 32). With respect to *2-sentence* outputs, 14 out of 16 were form 001 (*so*) outputs.

The results of the recall study must be tempered by the fact that a small sample (N = 8) was used. Even so, the findings shed considerable light on the ambiguous results of the comprehension experiment on causal relations (1.1).

First of all, it is clear that if there was a *cued* input, a *cued* output resulted. More importantly, if there was an *uncued* input, chances are there was still a *cued* output. It is possible that children store these relations in long term memory in a *cued* form regardless of the input they get. Furthermore, if they don't recall them in a *cued* form, the chances are at least 50-50 that they won't recall them at all. Given an *uncued* input, the frequency of errors was greater than the frequency of *uncued* outputs (6 vs. 5).

This helps to explain the fact that in the comprehension study there were no differences between forms in the number of correct responses made. The subjects apparently provided the necessary cues even when they were not there.

Another important finding in the recall study relates to effect of the *so* cue. In the comprehension study, the nature of the *why* question preempted a response beginning with *so*; there was a consequent reduction in the number of subordinated responses. The recall study provides a fairer test of the influence of this cue. The nature of the recall task does not stack the cards against the *so* cue by preempting a *so* response; it is reasonable to assume that a *so* response is equally as likely as a *because* response or an *uncued* response. The fact that other *cued* and *uncued* inputs were recalled as *so* inputs, indicates that it is a useful device for storing and

Table 10. Aided Recall of Causal Relations: Experiment 3—Input-Output Matrix

			Form of the Response								
		Code	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**	**Error**
Form of the Stimulus	**Because**	000	6	2	—	—	—	—	—	—	—
	So	001	2	6	—	—	—	—	—	—	—
	Because	010	1	—	6	—	—	—	—	—	1
	This was because	011	1	—	5	—	—	—	—	—	2
	and	100	2	2	—	—	1	1	—	—	2
	—	101	4	2	—	—	1	1	—	—	—
	and	110	2	1	3	—	—	—	1	—	1
	—	111	1	1	3	—	—	—	—	—	3
Total			19	19	17	0	2	2	1	0	9

recalling causal relations. In fact, it acted more like a subordinating form than a nonsubordinating form. Perhaps the view of traditional grammar (that *so* is a coordinating rather than a subordinating conjunction) is merely a grammatical convention. Psychologically it appears to be a subordinating conjunction.

One comment is necessary about the form 011 (*this was because*). The fact that there were no 011 outputs, not even from 011 inputs, indicates that the *this was* portion of the cue is superfluous. In fact, five of the eight 011 inputs were recalled as 010 outputs. The only difference between 010 and 011 is the insertion of a period and the inclusion of *this was*. Apparently that portion of the cue serves no function in memory and recall of causal relations.

The main conclusion to be drawn from the recall study is that, in order to store a causal relation, a subject virtually cannot help but store it in a unified, subordinated chunk. If he doesn't, he is just as likely to forget as he is to remember it. This finding provides strong evidence supporting the chunk model as a model of the psychology involved in processing verbal data. At least with respect to causality, people store data in unified rather than discrete units.

The small number of subjects used in the causal-recall experiment demands that it be replicated before too much credence is given to the results. It does, however, suggest an interesting hypothesis concerning children's strategies for storing semantic relations.

As an aside, one of the interesting outcomes from the aided recall study stemmed from the "dummy" items. "Dummy" statements (26) and (27) were often recalled in a form like (28), whereas "dummy" sentence (29) was seldom, if ever, recalled by any of the subjects. It is almost as if adjectival relations could be scaled on a compellingness-arbitrariness continuum. Arbitrary relations seem to serve an identification function, as in (29). Compelling relations seem to signal a more significant relation between the adjective and the noun, as in (27). The tallness of the man in (29) has little to do with the activity in the sentence; however, the anger of the bear in (27) is related to the activity.

(26) The bear chased the man. The bear was angry. The man was frightened.
(27) The angry bear chased the frightened man.
(28) The bear chased the man because he (or it) was angry.
(29) The tall man met the short woman.

In this connection, a study conducted by Myers (1967) provides confirming evidence. Myers asked subjects to choose paraphrases for sentences like:

(30) The slavegirl loved the kind master.
(31) The slavegirl loved the old master.
(32) The slavegirl loved the master. The master was kind.
(33) The slavegirl loved the master. The master was old.

When the elements within a sentence were highly related, as they are in (30) and (32), subjects tended to choose a paraphrase written in subordinated form, as in (34).

(34) The slavegirl loved the master because the master (he) was kind.

However, when the elements were not highly related, as in (31) and (33), subjects tended to choose adjective or coordinated paraphrases (and), like (31) or (35).

(35) The slavegirl loved the master, and the master was old.

All the causal sentences used in this study represented highly related units; all the effects were reasonable outcomes of the causes. Hence the tendency for nonsubordinated forms to be recalled in subordinated forms might well have been predicted from Myers' findings.

GENERAL DISCUSSION

With respect to the 3 theoretical positions outlined earlier—the readability hypothesis, the deep structure model, and the chunk model—the evidence favors the chunk model. Most of the data in the various experiments can be explained by the chunk model, while virtually none can be explained by the deep structure model or the readability hypothesis. This does not mean that the latter 2 positions are totally invalid, nor does it mean that the chunk model is a fully articulated psychological theory. The most reasonable conclusion is that, in general, any psychological model which attempts to explain the way in which verbal data are processed must begin with a semantic representation of the total relations involved rather than a syntactic description of the units which make up the relations. In short, some content must be put into the head before syntactic processing can occur. In this light, it will be interesting to examine the follow-up work of Bransford and Franks (1971) who seem to be attempting to develop a complete theory, accounting for perceptual as well as verbal phenomena.

The failure of the deep structure model to explain the present data says nothing at all about the validity of transformational-generative grammars as devices for representing the grammatical relations which occur in the language. There is no need whatever for a competence model (a model of an idealized speaker-hearer) to make any claims about a performance model (a model of real speakers, hearers, or readers). While it has been the hope of linguists and psychologists that a competence model would provide insight into the actual performance of language users, its failure to provide that insight does not, in principle, reduce its validity as a competence model. In short, a transformational grammar can serve as a powerful tool for generating all of the grammatical, but none of the ungrammatical, sentences of the language without making any claims about performance.

IMPLICATIONS

Pedagogically, the data lend no support to the recommendation that the difficulty of written discourse can be reduced by eliminating subordinating constructions or reducing sentence length. When the semantic relation is held constant and when the test question is relevant to the relation whose form is varied, either comprehension is equally efficient across forms or else the more subordinated and longer sentence forms elicit better comprehension. The well documented correlation between sentence length and comprehension difficulty should be viewed in one of 2 ways: a) the relationship exists because longer sentences are communicating more complex semantic relations than shorter sentences, or b) the relationship is an artifact of test questions which have measured semantic relations whose form is constant across longer and shorter versions of a unit of discourse.

While any recommendations concerning the most efficient surface structure forms for presenting various semantic relations must be tempered by the limitations of the present sample and testing procedures, the present findings certainly support an easing of concern for sentence length and complexity in the middle grades. However, before more specific recommendations can be made regarding structural efficiency, research is needed to assess the influences of the variables used in this study when sentences occur in naturalistic passage contexts. But the fact remains that, at least in this study, children seem not only to be able to handle complexity, but to actually prefer it.

The fact that grammatically more complex or longer statements equal or outperform their simpler or shorter counterparts is not surprising. What happens can be explained as a tradeoff relationship between explicitness on one hand and simplicity on the other. The causal relationship in (36) is explicit. If one rewrites (36) as (37), he has reduced grammatical complexity and average sentence length, but he has placed a new inferential burden on the reader.

(36) Because the chain broke, the machine stopped.
(37) The chain broke. The machine stopped.

What was previously complex but explicit becomes simple but implicit. A similar analysis accounts for the subjects' performance on adjectival relations. One limitation of the present study in this regard is that the statements selected for inclusion in the causal experiments all represented quite natural and predictable causal relations within a child's experience. One wonders whether or not children would be able to infer any causal relation from statements like (38), where their experience would be less helpful in making the inference.

(38) The new king clamped down on public meetings. Many residents emigrated to a new land.

The possible implications for social science and science content, where the intent is often to present *new* causal relations, are quite serious.

FURTHER RESEARCH

There are several follow-up studies suggested by findings in this study. It would be enlightening, for example, to see what happens to comprehension when statements similar to those selected for the present study are placed in paragraph or passage contexts. It may well be that when a subject is confronted with a search task in addition to a comprehension task, errors due to form are accentuated. The preliminary evidence from the Bormuth et al. study (Bormuth, Manning, Carr, and Pearson, 1970) would suggest that errors increase in contextual settings.

Second, the methodology used in this study could be applied to other classes of semantic relations, e.g., relations of adverseness (*however, although, in spite of*), time (*before, after, during,* etc.), purposiveness (*so that, in order to*), and conditionality (*if, once, provided that, unless*). This needs to be done to determine whether or not the conclusions regarding semantic cohesiveness are generalizable across a variety of verbal stimuli.

Third, traditional notions about readability should be reexamined. While the present study in no way suggests that readability formulas ought not to be used to predict the difficulty readers may encounter with particular passages, chapters, or books, it does suggest a new direction in research on readability. Studies need to be conducted in which different versions of a passage are constructed according to some rule-governed procedures, rather than according to the intuitions of the investigators. Furthermore, the questions used must be relevant to the structural changes which have been effected. In short, readability studies must begin with the question: What is the best way to communicate a given idea?

References

Bormuth, John R. Readability: a new approach. *Reading Research Quarterly,* Spring 1966, 1, 79–132.

Bormuth, John R.; Manning, John C.; Carr, Julian W.; & Pearson, P. David. Children's comprehension of between- and within-sentence syntactic structures. *Journal of Educational Psychology*, October 1970, 61, 349–357.

Bransford, John, & Franks, Jeffrey. The abstraction of linguistic ideas. *Cognitive Psychology,* October 1971, 2, 331–350.

Brown, James I. The Flesch formula. *Through the looking glass. College English*, April 1952, 7, 393–394.

Coleman, E. B. The comprehensibility of several grammatical transformations. *Journal of Applied Psychology,* June 1964, 48, 186–190.

Coleman, E. B. Learning of prose written in four grammatical transformations. *Journal of Applied Psychology*, October 1965, 49, 332–341.

Flesch, Rudolf F. The science of making sense. *American Mercury*, 1945, 60, 194–197.

Flesch, Rudolf F. *The art of plain talk*. New York: Harper and Brothers, 1946. (a)

Flesch, Rudolf F. How to say what you mean. *Science Digest*, 1946, 20, 37–39. (b)

Flesch, Rudolf F. *The AP writing handbook*. 1951.

Fodor, J., & Garrett, M. Some reflections on competence and performance. In J. Lyons & R. J. Wales (Eds.) *Psycholinguistic papers*. Edinburgh: Edinburgh University Press, 1966.

Gibbons, Helen D. Reading and sentence elements. *Elementary English Review*, February 1941, 18, 42–46.

Gough, Phillip B. Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior,* 1965, 4, 107–111.

Gray, W. S., & Leary, Bernice. *What makes a book readable*. Chicago: University of Chicago Press, 1935.

Hites, R. W. The relation of readability and format to retention in communication. Unpublished doctoral dissertation, Ohio State University, 1950. Cited by G. R. Klare, *The measurement of readability.* Ames, Iowa: The Iowa State University Press, 1963.

Holland, B. F. The effect of length and structure of sentences on the silent reading process. Paper delivered at American Psychological Association annual meeting. Chicago, 1933.

Klare, George R.; Shuford, Emir H.; & Nichols, William H. The relationship of style difficulty, practice, and ability to efficiency of reading and to retention. *Journal of Applied Psychology*, August 1957, 41, 222–226.

Klare, George R. *The measurement of readability.* Ames, Iowa: The Iowa State University Press, 1963.

Kueneman, H. A study of the effect of vocabulary changes on reading comprehension in a single field. Unpublished master's thesis. State University of Iowa, 1931. Cited by G. R. Klare, *The measurement of readability*. Ames, Iowa: The Iowa State University Press, 1963.

Marshall, J. S. The relationship between readability and comprehension of high school physics textbooks. *Dissertation Abstracts*, 1957, 17, 64. (Abstract)

Mehler, Jacques. Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, November 1963, 4, 748–762.

Miller, George A. Some psychological studies of grammar. *American Psychologist*, November 1962, 11, 748–762.

Miller, George A., Galanter, E.; & Pribram, K. *Plans and the structure of behavior.* New York: Holt, 1960.

Miller, George A., & McKean, Kathryn. A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology,* November 1964, 16, Part 4, 297–308.

Myers, William A. An experimental study of syntactic and semantic interaction. Unpublished doctoral dissertation, University of Minnesota, 1967.

Nolte, Karl F. Simplification of vocabulary and comprehension in reading. *Elementary English Review*, April 1937, 4, 119–124.

Orndorff, B. A. An experiment to show the effect of sentence length upon comprehension. Unpublished master's thesis, State University of Iowa, 1925. Cited by G. R. Klare, *The measurement of readability.* Ames, Iowa: The Iowa State University Press, 1963.

Savin, Harris B. Grammatical structure and the immediate recall of English sentences: 2. Embedded clauses. Cited in J. Lyons & R. J. Wales (Eds.) *Psycholinguistics papers.* Edinburgh: Edinburgh University Press, 1966.

Savin, Harris B., & Perchonock, Ellen. Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, October 1965, 5, 348–353.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw Hill, 1962.

Yngve, Victor H. Computer programs for translation. *Scientific American*, June 1962, 206, 68–76.

Yngve, Victor H. A model and hypothesis for language structure. Proceedings of the American Philosophical Association, 1960, 404, 444–446.

P. David Pearson is currently a professor of language, literacy, and culture in the Graduate School of Education at the University of California, Berkeley. Professor Pearson can be reached at ppearson@berkeley.edu.