GRAMMATICAL IMPLICATION AND ANAPHORA AS DETERMINANTS OF TEXT COMPREHENSION

(English as a Second Language)

by

DANIEL MICHAEL WALSH

B.A., University of Hawaii, 1967

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

Department of English Education

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1978

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of English Education

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date September 12, 1978

ABSTRACT

The present study seeks to assess the contribution of two aspects of linguistic competence, recognition of grammatical implication and identification of anaphoric referents, to a precisely defined measure in accord with a model that sets certain requirements for acts of text comprehension.

The first part of the research instrument presents items containing one or more sentences in isolation to test understanding of twelve syntactic transformations and four types of anaphora. The second part examines comprehension of sentences within a continuous text. The criterion measure for text comprehension is defined by two equally weighted components: indication of textual locus and judgment of truth value.

Sample populations include native speakers of English, Chinese speakers, and other non-native speakers learning English as a second language (E.S.L.). In each category, students at elementary, junior secondary, and senior secondary grade intervals are compared.

Analysis of the data reveals a developmental trend toward augmented performance on all variables. Native speakers as a group significantly outperform E.S.L. students only on the measure of recognizing grammatical implication relations which, in conjunction with the task of identifying anaphoric referents, contributes a greater proportion of variance to criterion scores than is observed in the case of either E.S.L. sample. It is concluded that native speakers and E.S.L. students attain equal proficiency in text comprehension by means of different strategies.

A comparison of the rank order of difficulty among the sixteen transformational and anaphoric types indicates that the scores of a multiethnic sample of E.S.L. students more closely approximate the pattern set by native speakers than do those of the Chinese speaking group, indicating that native language may influence performance on this sort of task. The rankings for all three groups generally support the notion that transformationally simpler structures are more easily interpreted.

Regarding the components of text comprehension, native speakers perform significantly better on the task of judging the truth value of an item statement than on indicating the locus of relevant statements in the accompanying passage. E.S.L. students encounter similar difficulty with both tasks. The locus indication component contributes somewhat more variance to the criterion scores of native speakers and multiethnic E.S.L. students than to those of the Chinese speakers. It is suggested that locus indication in itself may be a practical and reliable measure of text comprehension.

Some directions are offered for further research, recommendations are made for testing syntactic comprehension, and possible instructional implications arising from this and other research are discussed.

TABLE OF CONTENTS

Chapter

List of Tables

Acknowledgments

I wish to convey my sincere appreciation to those who contributed to the completion of the research and the preparation of this report, the limitations of which I am solely responsible.

Dr. B. A. Mohan for his continuing encouragement and receptive guidance in providing ample opportunity to explore the research topic in a manner that advanced my professional and educational growth

N. M. Ashworth for arousing my interest and confidence to pursue studies leading to this research which has been enriched by her perspective of practical applications and social value

Dr. K. Reeder for constructive insights that broadened the scope of the study

Dr. C. Pennock and J. Booth for offering perspectives from research in reading education

A. Moodie, Research Studies and Testing, Vancouver School Board, for prompt assistance in arranging access to suitable subjects

Teachers for their generous cooperation in administering the research instrument at a busy time of the school year

Dr. T. Rogers for extensive advice on the statistical analysis of the data

F. Ho and S. Kita for assistance in accessing the computer programmes

R. Berwick for advice on interpretation of some results and preparation of the manuscript as well as carefully executed graphics

T. Komori for assistance in preparation of the data

M. Seesahai for a careful proofreading of a lengthy manuscript

M. Bjornson and S. Haney for enduring support and assistance in my teaching assignment without which this research could not have been sustained

Chapter I

INTRODUCTION

Purpose of the study

The purpose of this study is to examine comprehension of a text in terms of one's ability to recognize grammatical implication. Specifically, such ability is manifest when one correctly judges whether or not a syntactic transformation is semantically equivalent to a given sentence or when one correctly judges whether or not the propositional content of a sentence can be deduced from that of a preceding sentence. In some contexts, either of these operations may require one to accurately identify anaphoric referents.

The present investigation is intended to assess the specific contribution of two linguistic skills -- (1) recognition of semantic equivalence between syntactic transformations or a grammatical implication relation from one sentence to another, and (2) identification of anaphoric referents -- to a criterion measure of text comprehension.

The study also seeks to determine whether the same order of relative difficulty among designated transformations and anaphoric types prevails for younger and older students and for native speakers and those who have acquired English as a second language.

Statement of the problem

What one perceives in attending to a text is the surface structure of sentences. To comprehend a sentence one must understand, at the very least, the relation of any logical subject and logical object (arguments)

to a predicate. Such underlying relationships are often expressed only in the deep structue of a sentence (cf. Chomsky, 1965). Hence, the comprehension process may be described as an act of deep structure recovery.

It is only be recovering the deep structure that one is able to recognize implications which are dependent upon the propositions that arise from the grammatical (argument - predicate) relations of elements within a sentence.

Deep structure recovery is therefore a requisite for comprehension of sentences and consequent recognition of grammatically based implication relations. It can be further claimed that recognition of implication is necessary for text comprehension.

Typically, tests of text comprehension seek to assess one's ability to derive valid conclusions which are, in fact, implications of stated propositions of sentences within the text.

In non-constructed response formats, one is confronted with a statement (e.g., Jack was told to hurry.), the truth value of which must be evaluated in light of the propositional content of earlier sentences (e.g., The woman called to Jack and Jill, "Hurry!"). This task clearly necessitates the recognition of grammatically based implication in order to compare the deep structures of two sentences that differ in surface structure.

Given this view of comprehension of a text as recognizing grammatical implication through deep structure recovery, to what extent do the linguistic skills of recognizing grammatical implication and identifying anaphoric referents contribute to performance, for instance, on a test of reading comprehension in which each item is transformationally linked to a specific sentence, or sentences, in the accompanying passage? The format proposed requires that one judge a statement to be semantically consistent,

contradictory, or indeterminate to the truth value of the related sentence(s) in the text. It will be assumed throughout that one's failure to understand a sentence results from difficulties of language comprehension in general rather than of reading comprehension in particular.

In appraising the contribution of the designated linguistic skills -- grammatical implication and anaphora -- to the proposed measure of text comprehension, two questions of practical interest might be addressed: Are students who have undertaken an E.S.L. programme in the elementary or secondary school as proficient in these linguistic skills as their native speaker peers? If not, does this place the E.S.L. student at a disadvantage on tests of text comprehension of the sort proposed to reflect deep structure recovery?

Further, by systematically defining a set of transformational and anaphora recognition tasks, it is possible to establish a hierarchical order of difficulty among the forms selected. On the basis of empirical findings, can such a hierarchy be accounted for by reference to transformational complexity? Are there notable differences in prevailing hierarchies for native speakers and E.S.L. students? Can these differences be explained in terms of current curricula in the teaching of English as a second language?

The research described in this report is intended to provide some tentative answers to these questions -- findings which may inform the design of more effective programmes to develop text comprehension.

Importance of the investigation

It is generally recognized that the ability to comprehend discourse is crucial to academic success. While, for example, innumerable components of the reading process have been proposed and examined at length, it is only

recently that educators have begun to focus upon the role of syntax as a significant factor in determining comprehension difficulty. Studies in this area should establish the fundamental importance of recognition of grammatical implication and anaphora relations in the overall comprehension strategy.

If syntactic complexity limits the native speaker's ability to comprehend a sentence, it is likely that the problem is more acute for the E.S.L. student. In the context of immigrant education, there arises a grave moral issue whereby the realization of social and economic aspirations of ethnic minorities is largely dependent on the ability to understand discourse containing in its structure logical relationships -- implications, presuppositions, entailments -- which the native speaker by virtue of his language competence comprehends intuitively. Adequate understanding of the content of discourse is also consequential to the foreign student whose career goals demand that he read professional literature published in English and be able to use the language to confer with native speaker colleagues.

Both the immigrant and the foreign student may be hindered from attaining some important text comprehension skills through a language teaching methodology that reflects the concentrated efforts of earlier linguists and language teachers to bring the student to greater levels of fluent performance. Worthy as this objective may be, such an approach generally fails to help the student appreciate how numerous syntactic patterns presented in isolation for mastery are semantically related to each other. It has been asserted in the opening statement that this knowledge is necessary for deep structure recovery, is certainly a requisite for recognition of grammatical implication, and cannot be ignored when

analyzing the act of comprehending a text. More pragmatic conceptualizations of second language learning will need to address this problem by focusing student attention on the semantic interpretations of syntactic structures.

Teachers who have come to appreciate the importance of syntactic - semantic relations to comprehension have further reason to be distressed by the state of current testing procedures that seldom define which grammatical relations are being examined. Lacking this definition, ordinary comprehension tests must be seen as collections of items whose definitions are rooted entirely in the mind of the test author and, therefore, not objectively verifiable. Such tests are useless to a teacher who wants to know which sorts of syntactically based relations his student has failed to comprehend, information essential for remedying deficiencies in the fundamental skill of sentence comprehension. One aim of the present study is to extend earlier research in developing diagnostic instruments that promote purposeful instruction.

In arguing the importance of a heretofore neglected element of English language learning, this study cites weaknesses in instructional and testing practices. Through its examination of the contribution of certain linguistic skills, a better understanding of some of the components of comprehension may be achieved. This knowledge, based upon data from E.S.L. students and native speakers, may provide insights that could eventually improve the quality of instruction available to both.

Overview of the study

In a review of the literature, Chapter II, evidence for the relationship between syntax and reading comprehension is cited from a number of different perspectives. This is followed in Chapter III by a description

of each facet of the present instrument drawing comparisons with previous attempts to measure understanding of syntactic and anaphoric relations. Requirements for comprehension testing are proposed through a model which seeks to rule which cognitive acts are indicative of comprehending the text at hand.

In Chapter IV, the research design is presented and the hypotheses stated. A rationale is offered for testing each hypothesis. Chapter V outlines the procedures taken to assemble the tests, select suitable Ss, administer and score the instrument.

After reviewing the statistical properties of the instrument, Chapter VI examines each hypothesis, draws a conclusion, and briefly discusses the possible causes and practical implications of the findings.

Results of the study are summarized in the final Chapter VII where are noted directions for further research, recommendations for alternative testing procedures, and some implications for second language instruction.

Limitations of the study

Circumstances in the research setting forced the investigator to rely largely upon teacher report for the selection of suitable Ss. The specific criteria employed are outlined in the section describing sample populations. Information regarding I. Q. scores, measured achievement in reading comprehension and other English language skills, and records of past performance in other academic areas is not usually available to researchers. To administer such measures in addition to the study instrument would have placed an unreasonable demand upon the time of the participants at the close of the school year.

Another difficulty encountered by the investigator was that of access to an ample number of Ss in all three native language categories. As a result, the sample size of native speakers of English at the elementary and junior secondary intervals and of non-Chinese E.S.L. learners at all grade intervals is admittedly small.

Attention is given to comparing Ss with regard to the factors of native language and grade placement. One expectation is that native speakers will perform superiorly to E.S.L. students on all tasks. Another is that both first and second language learners will exhibit a pattern of augmented scores as a function of higher grade placement. The latter incurs a problem of valid comparison across grade intervals. Ideally, a longitudinal study is best suited to investigating any pattern of growth. The present study uses a cross-sectional model which may be valid in the case of native speakers as it may be assumed that, in general, Ss at higher grade intervals possessed, at an earlier point in their language development, a constellation of linguistic skills approximating that of Ss at lower grade intervals. On this basis, one may draw inferences from the data regarding the gradient of learning between those grade intervals selected and defined for this study.

The foregoing assumption, however, becomes untenable when applied to an E.S.L. population. Typically, immigrant students enter the host school system at various ages, spend approximately one year in a special language training class, and are then placed in regular classes with their peers. It is doubtful that all immigrant students gain a similar configuration of language learning through this sort of experience. Another factor accounting for variation in learning among any sample of immigrants is the extent and nature of the individual's educational experience in his country of origin which may, or may not, have included English

as either a subject or medium of instruction. Previous schooling of students from diverse backgrounds is difficult to assess in constant and objective terms and therefore could not be controlled in this study. Since the majority of E.S.L. participants had resided in Canada for less than two years, it cannot be said that Ss at higher grade intervals would have been accepted as suitable Ss for lower grade intervals at a point earlier in their chronological growth. Consequently, one must be wary of any attempt to draw conclusions about the gradients of change in performance across grade intervals within a non-native population.

A final limitation to be kept in mind arises from the practical need to restrict the number of transformations and anaphoric types to be tested and to select a particular format to test text comprehension. Conclusions can be based only upon the relationship between facility in the tasks included in the present instrument and performance on text comprehension tasks of a specified nature which may depart from other, more conventional, means of measuring text comprehension.

Chapter II

REVIEW OF THE LITERATURE

In this chapter, a number of well established principles are recalled in order to place the present study in the context of recent research in the area of syntactic comprehension as related to readability and growth in reading proficiency. Unlike the present study whose domain is language comprehension in general, sources cited herein are particularly concerned with reading comprehension. Nevertheless, it is considered that many insights gained from these investigations can be generalized to comprehension of both written and spoken texts. Further applicability lies in the fact that the problem at hand is examined through the use of printed materials. How the results of the present study relate to certain of these earlier investigations will be outlined in the final chapter of this report.

Sentence comprehension is something more than word comprehension.

Psycholinguists have given considerable attention to the fundamental role of syntax in comprehension.

> Perhaps the most obvious thing we can say about the significance of a sentence is that it is not given as the linear sum of the significance of the words that comprise it....the words in a sentence interact. (Miller, 1965:16)

> The rules for this interaction of words in sentences are set by the syntax of a given language; the outcome of this interaction is meaning. (Cooper and Petrosky, 1976:187)

Implications for deriving meaning from the reading process can be drawn from the foregoing propositions:

> Since psycholinguistics asserts that the goal of fluent reading is the identification of meaning... which relies heavily on the brain's ability to bridge surface structure and deep structure with syntactical rules, it is obvious that great weight in thinking about the reading process must be put on the relationship of words in sentences. (Cooper and Petrosky, ibid.)

Empirical support for this point of view is presented in an early study by Gibbon (1941) which, through the use of a "disarranged phrase test", established among a Grade 3 population a high correlation (.89) between the ability to perceive relationships between parts of a sentence and the ability to understand the sentence, when intelligence is partialled out. Also, a significant correlation was demonstrated between the ability to see syntactic relationships and total reading achievement. More recent investigations conducted by Cromer (1970); Oakan, Wiener, and Cromer (1971); and Steiner, Wiener, and Cromer (1971) indicate that the poor comprehension of some readers is not due to weak skills in word identification but to an inability to integrate the meanings of separate words to arrive at the meaning of an entire sentence.

The relevance of these empirical observations, particularly to the instruction of E.S.L. students, should be clear:

> A method of teaching reading which stops with recognition of words...assumes that the pupils have acquired the ability to supply the proper grammatical components of meaning as they have learned to speak the language. When children are already fluent speakers of English and are familiar with the standard usages, word recognition may be enough...Something more than word recognition is indicated...when children progress to written material which is structurally different from conversational English. (Ives, 1964:180)

To develop the reading comprehension of the E.S.L. student whose exposure to English is primarily in the domain of peer interaction and/or instructional materials that emphasize the spoken language, teachers must come to recognize the crucial role of syntax in deriving accurate meaning from written discourse as well.

Syntactic units are used to organize sentence perception.

Miller and Isard (1963) demonstrated that words in grammatical sentences are easier to perceive aurally than words connected in ungrammatical strings. Morton (1964) demonstrated a steady increase in speed of reading as a series of words approximates normal syntactic patterns. Epstein (1961) showed that "syntactic structure facilitates verbal learning apart from the contributions of meaningfulness, familiarity, and sequential probability" (p. 85). Wisher (1976) discovered that beforehand knowledge of the syntactic structure to be used facilitates performance on a memory task and decreases the time required to read the sentence. Kolers (1970) reported that 70% of the oral reading substitutions of adults conform to the same part of speech as the correct word in the text. This is indicative of an intuitive awareness by the reader of syntactic constraints.

In the context of learning to read, Goodman and Burke (1969) have noted that more than 60% of the oral miscues of elementary school children observed could be described as retransformations of the text sentence rather than simply anomalous strings. A study by Weber (1970) among Grade 1 pupils recorded that 90% of the oral reading errors did not contain syntactic violations, suggesting that beginning readers may be over reliant on their still limited knowledge of syntax. Further evidence of the influence of syntax on reading behaviour is offered by Rode (1974) who showed that the

eye-voice span -- the number of words that the eye is ahead of the voice in oral reading -- among pupils in Grades 3, 4, and 5 is in some ways determined by syntactic structure.

Syntactic comprehension is characterized by developmental trends.

The extensive research of O'Donnell, Griffin, and Norris (1967) affirms the general notion of a developmental sequence of syntactic acquisition throughout the elementary grades. These investigators found that some transformations (e.g., relative clauses) were used much more frequently by kindergarten youngsters while other structures (e.g., noun modification by a participle) appear in the language of older children. Ruddell (1969:11) interprets such findings as logical from the standpoint of transformational-generative grammar in that many of the later constructions are derived from more complex deletion rules.

Marcus (1971) devised "A Test of Sentence Meaning" for Grades 5 through 8. Results show a pattern of improved performance as a function of grade level. Such findings are consistent with those reported by Carroll (1970), Smith (1970), and Tatham (1970).

School aged children are unable to comprehend many syntactic structures.

Bormuth et al. (1970), having tested more than 240 Grade 4 students' comprehension of a wide variety of sentence structures (including nominalization, relativization, subordination, and anaphora) concluded that their sample population "showed an unexpectedly low level of performance on these skills which seemed both very simple and very basic" (p. 349). Marcus (1971) and Takahashi (1975) have demonstrated that among students in upper elementary

and junior secondary grades, no group as a whole appears to have completely mastered an understanding of the transformational types included in "A Test of Sentence Meaning"; pronominal reference, deletions, embeddings, and conjunctions are cited as particular sources of difficulty . In a more specialized study, Stoodt (1970) also found Grade 4 students frequently encountered difficulty interpreting sentences containing certain conjunctions.

All of the foregoing studies generally confirm the earlier observations of Chomsky (1969) who determined that, at a given age, not all children are able to demonstrate the same level of mastery of selected syntactic structures.

Some syntactic patterns are easier to comprehend than others.

Research in this area is extensive and requires careful analysis and interpretation.

For developmental and sociolinguistic reasons, the syntax of oral language can be considered to be both more familiar and transformationally simpler than many of the patterns encountered in written discourse. Therefore, many investigators have attempted to relate aspects of oral syntax to reading comprehension. Tatham (1970) concluded that Grade 2 and Grade 4 children are better able to comprehend syntactic patterns in reading if those patterns are frequently used in their oral language. Ruddell (1969) compared the effect on reading comprehension of written patterns of language structure which occur with high and low frequency in children's oral language. Among a sample of Grade 4 pupils, reading comprehension scores on passages written with high frequency patterns were found to be significantly superior to comprehension scores on passages that contained low frequency patterns. Similarly, Reid (1972) rewrote basal reader sentences to match

the spoken syntax of her 7 and 8 year old Ss. A comparison of comprehension between the original and the rewritten material revealed the latter to be significantly easier for these children to understand. Finally, Vogel (1975) conclusively demonstrated that dyslexic children are frequently deficient in oral language syntax.

The facilitating effects of oral syntax well established, there remains the question of why this is so. Evidence can be found in the literature to support the case for either transformational simplicity or familiarity.

A touchstone study by Coleman (1964) established that reading a "detransformed" text in which nominalizations, passivizations, relative clauses, and grammatical deletions were not present resulted in significantly higher cloze test scores among a group of 48 college students. Fagan (1969) also observed significantly higher cloze test performance when elementary students in Grades 4, 5, and 6 read texts that did not contain relative clauses or grammatical deletions. Evans (1973) demonstrated the superior comprehension of simplified (de-transformed) prose on multiple choice tests as well as cloze measures. It is worth noting, however, that the Grade 12 Ss in this study were previously identified as reading at three to five years below grade level. (Goodman and Burke, 1973, claim that differences in ability to handle complex syntax disappear among readers of moderate to high proficiency. Evans recognizes the significance of this factor in hypothesizing that problem readers will raise their comprehension by reading transformationally simpler prose.) Dealing with a more specific problem in syntactic comprehension, Richek (1976) required pupils in Grades 3, 4, and 5 to identify subordinate clause subjects with two levels of MDP (minimal distance principle), conforming and violating, and complexity, following and

interrupting statements. That investigator reports: "Significant main effects were found for MDP (conforming sentences were easier) and complexity (following statements were easier). The MDP and complexity variables formed a significant interaction. The MDP violating sentences produced performance characteristic of short term memory tasks, making complex sentences which separate subject and subordinate clause by several words, difficult to process." (p. 800).

Results of the four preceding studies support the case for transformational simplicity as a key factor in comprehension. More problematic is an investigation by Peltz (1974) which required 34 Grade 10 Ss to write a page on social studies content. Their writing was compared with the syntax of their textbook and a text passage was rewritten to conform to the students' syntax. Ss' comprehension of the original and the "simplified" passages were then compared. Findings indicate a significant difference in favour of the simplified version when comprehension was measured by means of a cloze test. Use of a multiple choice format, however, showed no significant difference in comprehension between the original and rewritten versions. Reviewing the methodology used in this study, it is debatable whether one version may have been easier because its grammatical transformations were simpler or because its syntax was more familiar to the Ss.

At least two other studies, however, tend to support the case for familiarity of syntactic structure effecting better comprehension. Smith (1970) presented students in Grades 4 through 12 with four cloze tests exhibiting syntactic characteristics of the written productions of students in Grades 4, 8, and 12 and skilled adults. Results indicated that students in Grades 4, 5, and 6 read Grade 4 syntax best; Grade 11 students read it with least facility. Students in Grades 8 through 12 found Grade 8 syntactic

patterns easier to read than either Grade 4 writing or the more difficult passages. Pearson (1974) gathered evidence from the performance of children in Grades 3 and 4 which tends to refute 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 subordinate and longer sentence forms elicit better comprehension. (p. 189)

Further research seems necessary to determine more precisely how syntactic familiarity and complexity interact in comprehending a text and to define those conditions under which one factor becomes more prominent than the other.

Syntactic comprehension contributes to more generalized measures of reading comprehension.

Studies conducted at every grade level sustain this proposition. Harris (1975) established that for Grade 2 pupils performance on an oral syntax test showed a high correlation with scores on a standardized measure of reading achievement (.70). This correlation was significantly greater than the correlation between measures of reading achievement and intelligence (.57). The importance of syntactic attainment as a predictor of reading achievement at this age level has also been demonstrated by Vogel (1975) in which the proportion of unique variance contributed by nine measures of syntax to a criterion standardized test of reading achievement amounts to 53.1% for mormal (non-dyslexic) Ss, an impressive result considering the need of young readers to consciously attend to other tasks --

e.g., decoding, word recognition.

Stoodt (1970) observed that among Grade 4 students, there is a significant relationship between reading comprehension and understanding conjunctions. Hart (1971) administered a sentence combining test to Grade 6 pupils identified as reading at a Grade 3 level and determined that reading comprehension scores were related to the ability to produce sentences that carry the informational load of several kernel sentences. Similarly, Kuntz (1975) demonstrated a close relationship between Grade 7 students' performance on a written sentence construction test and their scores on a widely used standardized reading achievement test.

Differences between typical and debilitated readers with respect to syntactic capability have been established in a number of instances. Vogel (1975) has shown the syntactic comprehension of Grade 2 male dyslexics to be inferior to their normal achieving peers. Takahashi (1975) used "A Test of Sentence Meaning" (Marcus, 1971) to compare able and poor Grade 9 readers on their ability to interpret sentences. On the basis of a significant difference between these groups, she concluded that "comprehension of syntactic structure is an element in total reading comprehension." (p. 60) Cox (1976) developed a test of syntactic complexity and found that adult beginning readers enrolled in a "basic education" programme performed significantly inferiorly to literate adults. A study conducted by van Metre (1974) compared some of the linguistic competencies of bilingual Grade 3 pupils drawn from the top and bottom quartiles on a standardized reading achievement test with matched groups of monolinguals. In oral interviews, all Ss were tested for comprehension of four syntactic structures described by Chomsky (1969) -- ask (query) / tell, promise / tell, easy to see, and pronominalization. Findings revealed that greater differences occur between high and

low reading achievers than between bilinguals and monolinguals. These results are particularly interesting not only because they confirm the importance of syntactic comprehension for overall reading achievement but also for their practical implications in suggesting possible causes of a second language student's unsatisfactory progress in learning to read English.

Syntax is an important variable in determining readability.

The preceding arguments concerning the distinction between sentence comprehension and word comprehension and the organizing function of syntax have not historically influenced the construction of readability formulas. Botel and Granowsky (1972) note that sentence length is the only syntactic measure in many widely used formulas -- e.g., Dale-Chall (1948), Spache (1953) -- and argue that a syntactic analysis based on transformational grammar indicates the complexity of a sentence should not be judged solely on a word count of the surface structure of the sentence. By identifying specific linguistic variables, Bormuth (1966) has claimed to have accounted for a far greater proportion of variance in comprehension difficulty than was possible with earlier formulas. Granowsky (1973) also reports an application of transformational grammar theory to the development of a syntactic complexity formula which promises to be a more reliable and valid guide to determining readability.

Chapter III

DEVELOPMENT OF THE INSTRUMENT

This chapter describes the underlying linguistic concepts which form the basis for each task comprising the present instrument.

TESTING SYNTACTIC COMPREHENSION

The first part of the instrument examines understanding of syntactic relations within a single sentence or between two sentences isolated from any larger context.

Studies reviewed in the preceding chapter claim that skill in syntactic comprehension contributes to success in more generalized measures of reading comprehension. This aim raises the question of precisely how it can be demonstrated that one has understood a sentence.

In responding to comprehension test items, it has been suggested that the reader bases his guesses on as few lexical, structural, and graphic clues as possible, aided by the fact that language is redundant and sequential (Goodman and Burke, 1969). Hence, a pupil may generate or select a correct response coincidentally upon such minimal clues. More limited than the native speaker in his knowledge of the language, the E.S.L. student may be even more dependent on superficial clues from the surface of the text. A growing appreciation of this problem has given rise to new demands for operationally based tests which clearly demonstrate what linguistic properties are being tested (Bormuth, 1970; Mohan, 1973).

Grammatical implication

In the context of this study, the term "grammatical implication" is restricted to those instances wherein the propositional content of one sentence is implied by that of another, preceding, sentence by virtue of a reordering of syntactic surface elements. Deletion or insertion of elements may occur. Thus, recognition of implication may be viewed as a matter of accurately recovering the deep structure of each of two sentences which differ in their surface structure but have a defined transformational relationship. In the case of two-way implication relations the surface structure of both sentences reflects an identical deep structure, e.g.,

Harry ate the cake.
The cake was eaten by Harry.

The sentences in a one-way implication are characterized by both different surface structures and different deep structures. The implication relation rests on the deep structure of the second sentence being closely related to that of the first, e.g.,

Harry ate the cake.
The cake was eaten.

A native speaker recognizes that if the first sentence is true, then the second sentence is true, but not vice versa.

On the basis of the foregoing stipulations, given the sentence

(1) The black cat jumped over the fence.

only a syntactic paraphrase such as

(2) The cat which is black jumped over the fence.

is here considered a "grammatical implication" although it is equally necessary to recover the deep structure of

(3) The black feline jumped over the fence.

to recognize it to be a lexical paraphrase of (1) or

(4) The black cat went to the other side of the fence.

to judge it a logical implication of (1).

By limiting "grammatical implication" to syntactic paraphrases it becomes possible, through the classifications offered by transformational grammar, to define the linguistic properties any item purports to test.

Previous attempts to measure understanding of grammatical implication

The investigator examined three recent research instruments which relate to grammatical implication and may be considered as attempts to meet the requirements of stating what aspects of linguistic competence are to be tested. All claim their analysis to be based on transformational - generative grammar and seek in various ways to evidence the S's capability to determine the deep structure of a given sentence.

Simons (1970) devised the "Deep Structure Recovery Test". Twenty-five items test understanding of transformations mainly concerned with syntactic - semantic contrasts among a set of transitive and intransitive verbs and the ask / tell distinction (cf. Chomsky, 1969).

Ss were asked to indicate which of three sentences differs in meaning from the other two. The following is a sample item:

* a) What the boy would like is for the girl to leave.
b) For the boy to leave is what the girl would like.
c) What the girl would like is for the boy to leave.

The test was administered to 87 Grade 5 pupils with a mean I Q of 117. On the average, these students correctly answered approximately 75% of the items.

On the basis of a positive correlation between scores on this test and performance on a standardized test of reading achievement (.48) and cloze measures (.73), Simons concluded that the ability to recover the deep structure of sentences is an important aspect of reading comprehension.

Marcus (1971) constructed "A Test of Sentence Meaning" to determine intermediate grade students' understanding of syntactic clues to literal meaning (p. 50). While structuralist categories were used to isolate the types of syntactic structures to be included in the test, transformational theory was used in developing the items related to specific skills of interpretation. From categories of modification, predication, and coordination, seventeen syntactic structures were identified and six related items were presented for each.

A number of formats are used. In the first example, the student has to find the transformation that has the same meaning as the underlined sentence.

<u>The man gave the boy a puppy.</u>

a) The man gave away the boy's puppy.
* b) The man gave a puppy to the boy.
c) The boy gave a puppy to the man.
d) The man gave a puppy away for the boy.

An alternative format requires the student to select one of four sentences that utilizes the vocabulary of the other three but differs from them in meaning. Still other formats require the reader to analyze a given structure into its basic kernel sentences. In the following example, one is to choose the <u>two</u> sentences that combine to give the complete meaning of the underlined sentence.

Bob and Don ate the bread and jelly.

* a) Bob and Don ate the bread.

b) Bob ate the jelly.

c) Don ate the bread.

* d) Bob and Don ate the jelly.

e) Don ate the jelly.

The test was administered to 487 students in Grades 5 through 8 in both "disadvantaged" and middle class area schools. Results show a trend of improved performance as a function of higher grade level. Grade 5 pupils averaged 60% correct responses; Grade 8 students reached an average of 80%.

An analysis of errors reveals that "some students mistakenly thought that a coincidental noun - verb - noun sequence of words was a subject - verb - object sequence and thus a kernel sentence of the larger sentence." (p. 58). Such an error would be an instance of failing to recover the deep structure of either the original sentence or one or more of the options suggested to be equivalent in meaning. Another interesting observation, from the viewpoint of grammatical implication as defined in the present study, is that some students apparently "did not distinguish between denotated literal meanings and implied meanings." (ibid.)

No attempt was reported to correlate test results with performance on other measures of reading comprehension.

O'Donnell (1973) recognized the need for a test to measure awareness of syntactic structure without relying on the terminology of grammar. To this end, the "Perception of Alternate Structures Test" was devised using nonsense vocabulary to encourage reliance on syntactic, rather than lexical, cues to structure. According to the author:

> Of the thirty items on the test, two measure perception of the active - passive alternatives, two of the indirect object - prepositional phrase options, six the relative clause - reduced relative variations (prenominal adjectives, participle phrase and appositive) and two the adverbial clause - abridged adverbial alternatives. Six items deal with noun clauses - infinitive - gerund phrase variations, and the remainder of the items test various combinations of the options listed above. (pp. 3-4).

Each item contains three sentences, two of which are similar in underlying meaning. The student must indicate which sentence is different from the others. A sample item follows:

a) The birtle scared the ilbid.
b) The ilbid was scared by the birtle.
* c) The ilbid scared the birtle.

The test was administered to 87 Grade 9 students and 62 Grade 10 students, approximately half of whom scored below the 35%ile on the cognitive abilities (verbal) section of a widely used standardized test. On the average, Grade 9 students answered 44.3% of the items correctly, Grade 10 students, 50.5%.

The study did not establish any correlation between this test and other measures of reading comprehension.

The present instrument

Referred to elsewhere in this report as Task 1, this part of the present instrument invokes two distinct series of questions. Test No. 1 sets a task of syntactic paraphrase recognition similar to that required by the "Deep Structure Recovery Test" (DSRT) on which some items were modelled and the "Perception of Alternate Structure Test" (PAST) as well

as certain of the formats used in "A Test of Sentence Meaning" (ATSM). Following Mohan (1973:97), paraphrase is here defined as a two-way implication relation: "An assertion A_i is a paraphrase of A_j if A_i implies A_j and vice versa." Test No. 2 sets a task of recognition of one-way implication relations similar to that required by certain formats of ATSM. Unlike the latter, however, it does not directly show a recovery of underlying kernel sentences; rather, the S is merely asked to judge if A_i implies A_j.

In determining the optimal format for Tests No. 1 and No. 2 (q.v. Appendix "B"), it was considered preferable to require Ss to compare only two sentences at a time for semantic equivalence as defined above. By doing so, the memory burden, particularly for younger pupils, might be reduced thereby affording a truer measure of linguistic competence. For example, in an earlier study conducted by this investigator, the following item was adapted from the sample of the DSRT cited above:

> What the boy would like is for the girl to leave.
> For the boy to leave is what the girl would like. yes *no

96.3% of a sample of twenty-seven average Grade 6 students answered this item correctly in contrast to only 70.9% correct responses among Grade 5 pupils of superior intellectual ability to the original DSRT three sentence item.

To accomplish the purpose of the present investigation, certain other features of the aforementioned tests were examined and considered unsuitable. The uneven sampling of transformations in the DSRT and the PAST does not facilitate direct comparisons of difficulty among transformational types. The use of nonsense vocabulary in the PAST presents a

task which appears not only artificial but also may be not especially applicable to investigating the variables defined in the research problem as that test's author concedes:

> Those tests that utilize nonsense vocabulary to encourage reliance on syntactic cues have a low correlation with reading comprehension tests, while those that utilize conventional vocabulary and allow reliance on semantic as well as syntactic cues have noticeably higher correlations with reading. (O'Donnell, 1976:4).

ATSM, while based on a more comprehensive syntactic typology than either of the other two, does not test the transformations of greatest interest to this investigator. As well, multiple response formats may have proved too complicated for younger Ss. Lastly, the length of the test --- 102 four and five option items -- would have substantially increased the time required for administration since the research design sought to assess the contribution of the syntactic tests to measures of text comprehension.

This investigator therefore elected to construct for the first part of the instrument, a group of thirteen four-item subtests to appraise understanding of passivization, relativization, and transformations involving the "minimal distance principle" as well as paraphrases of indirect speech and pseudoimperatives.

Anaphora

Apart from the sentence transformations highlighted in the instruments described thus far, another syntactic paraphrase device -- anaphora -- is thought to be frequently associated with comprehension difficulties encountered by second language learners. Anaphora is the term used to denote a structure in a sentence -- e.g., a pronoun, a pro-verb, or a clause demonstrative -- that derives its meaning from another part of the present sentence or from another sentence, usually one that occurs previously in the text.

Previous attempts to measure understanding of anaphora

From a taxonomy of anaphoric structures developed by Bormuth (1970) and Menzel (1970), a test was devised by Bormuth et. al. (1970) to check comprehension of anaphoric relations. A multiple choice format was used. These investigators established a rank order of difficulty among fourteen anaphoric types for their sample of 240 Grade 4 students. Oddly, the easiest structure proved to be pro-clause forms, the most difficult, common personal pronouns.

Lesgold (1973) examined the comprehension of 80 Grade 3 and 4 children on fourteen varieties of anaphoric structures including several examined by Bormuth et. al. Using a "wh-- question, constructed response" format, the results obtained indicate a marked disparity between the two studies with respect to the ranking of those anaphoric structures tested by both investigators. It was found, for example, that the easiest of all forms tested in this later study were personal pronouns, the pro-clause being one of the most difficult.

Neither Bormuth et al. nor Lesgold report correlations between their anaphora tests and other measures of reading comprehension.

The present instrument

The present investigator defined four general categories of anaphora on the basis of the descriptive classification offered by Halliday and Hasan (1976) to which was added one type (Subtask R) adapted from Chomsky (1969). The predicted order from easiest to most difficult was:

Subtask P:	Pronominal reference	(7 items)
Subtask R:	Pronominal reference violating the minimal distance principle	(5 items)
Subtask S:	Nominal substitutes	(5 items)
Subtask T:	Clausal substitutes	(3 items)

A total of twenty items as indicated above were prepared for Test No. 3 using a "constructed response substitution" format (q.v. Appendix "B") to comprise this portion of the instrument, designated as Task 2.

TESTING TEXT COMPREHENSION

For the purposes of this study, text comprehension is intended to mean the recognition of test item statements as containing propositions consistent, contrary, or indeterminate to those overtly stated in one or adjacent sentences of the accompanying text.

Previous attempts to measure understanding of a text

Teachers and researchers will recognize that the foregoing objective appears to motivate many, but by no means all, of the items

included in the "paragraph comprehension" sections of standardized reading achievement tests and in the less formal inventories contained in published materials which are intended to develop comprehension of texts.

In reviewing past research literature, a number of studies were cited which claimed to demonstrate that syntactic understanding is crucial to reading comprehension. Usually, the criterion measure employed was some well known standardized test. Future investigations designed to test the relation between syntax and overall comprehension must invoke more rigorous definitions of comprehension than those operative in most published tests. The present study subscribes to a definition set forth by Bormuth:

> ...comprehension ability is thought to be a set of generalized knowledge - acquisition skills which permit people to acquire and exhibit information gained as a consequence of reading printed language. (1969:50).

If this definition is further restricted to include only exhibitions of information gained from a reading of the material immediately at hand and not from earlier reading experiences, the dubious validity of popular standardized tests of reading comprehension should become apparent.

> Comprehension is defined as the ability to acquire information from a passage, but one tries to measure it by finding out how many questions the person can answer on a test given him only after he has read the passage. This procedure ignores the facts that it is almost impossible to find a passage dealing with information about which a person knows

> absolutely nothing and that he could probably have used this information to answer some of the questions even before he had read the passage. (Bormuth, 1969:52).

Empirical support for this contention may be found in a number of studies on passage dependency. One of the most notable (Tuinman, 1973) examined five major tests of reading comprehension. Approximately 1,800 students in Grades 4, 5, and 6 participated in this study. Results indicated that none of these widely used tests provided sufficient guarantees against the examinee's answering items on the basis of information other than that presented in the passage. Average probabilities of correct responses, even when the passage was not present, were well above chance scores. Prior knowledge, elimination of irrelevant distractors, and the use of information embedded in preceding questions are all suggested as possible causes for unexpectedly high scores.

The present instrument

To cope with such problems in testing, attention must be given to an important distinction.

> ...scores on comprehension tests given in the usual way have two components: those questions the student could have answered without reading the passage and those questions he was able to answer only as a consequence of reading a passage. Only the latter may be definitely said to represent knowledge gained through reading. (Bormuth, 1969:52).

This clarification prompted the investigator to construct a model for testing comprehension of a written text, given certain qualifications, that seeks to discriminate between terminal behaviours that constitute acts of text comprehension and those that do not.

Having considered through the model (q.v. Appendix "A") a number of possible strategies that could lead to the selection of the correct response option, the essential problem in evaluating an examinee's set of responses is a matter of recognizing those legitimately derived from an immediate reading of the text and eliminating those correct responses attained through earlier recollections or false comprehension strategies. Given the proposed format and the stated qualifications, the model posits two distinguishable aspects of any act of reading comprehension -- (1) the actual location of the statement in the passage that is relevant to an interpretation of the item (or the capability to do so on demand) and (2) an accurate recovery of the deep structure of both the passage and the item statements in order to determine, if possible, the truth value of the item statement. Any assertion that a student has comprehended must be supported by a demonstration of skill in both aspects. Conventional tests, however, have ignored the possibility of asessing the first aspect, identification of textual locus, thereby relying solely upon an apparent indication of the second aspect, judgment of truth value. Comprehension tests constructed in this fashion are inadequate for the reason that while locus identification is a prerequisite to an act of text comprehension, it is not essential to the coincidental selection of the correct response to multiple choice items as studies on passage dependency have shown.

In order to make more direct observations of skill in both aspects, the investigator set two separate tasks for Test No. 4, the second part of the instrument specifically intended to test sentence comprehension within a written text. Twenty-three item statements relating to an accompanying passage were presented. In each instance, the S was asked to identify the numbered sentence(s), if any, in the passage judged to provide information appropriate to determining the truth value of the item statement. This locus indication task is referred to in this report as Task 3. The other task was to appraise the item statement's truth value through selection of one of three response options: "true", "false", or "cannot say". This latter task is referred to in this report as Task 4. The labelling of these two separate measures as Tasks 3 and 4 is not intended to suggest that the S always performs the locus operation prior to judging the truth value of an item statement although the actual format of the test might have invited him to do so. Depending on strength of memory, a S might more readily recall what had been stated in the passage than the point at which the relevant statement occurred. Since this factor could not be controlled by the investigator, Ss were not given specific direction to carry out one task before the other.

In the context of the present investigation, an act of text comprehension is considered to have occured only when a S offered a correct response to both the Task 3 (Locus) and the Task 4 (Truth Value) components of any given item. This combined performance is referred to in this report as Task 5 (Locus + Truth Value). Task 5 was not an additional operation for the Ss; rather, it is an essential construct

used by the investigator in scoring Test No. 4, the text comprehension part of the instrument. It is Task 5 against which Ss' performance on Task 1 (Grammatical Implication) and Task 2 (Anaphora) is compared.

Chapter IV

RESEARCH DESIGN AND HYPOTHESES

For preliminary analysis, this study employs a 3 x 3 factorial design to effect comparison among sample populations with respect to: (1) a language factor (N.S. = native speakers of English; E.S.L.(A) = English as a second language students who are native speakers of Chinese; E.S.L.(B) = English as a second language students who are native speakers of a language other than Chinese) and (2) a grade placement factor (Elem = Elementary, i.e., Grades 4, 5, and 6; JrSec = Junior Secondary, i.e., Grades 7, 8, and 9; SrSec = Senior Secondary, i.e., Grades 10, 11, and 12).

	Elem	JrSec	SrSec
N.S.			
E.S.L.(A)			
E.S.L.(B)			

Comparisons are conducted for each of five task variables: (1) recognition of grammatical implication; (2) identification of anaphoric referents; (3) indication of textual locus; (4) judgment of truth value; and (5) a criterion measure of comprehension based directly on performance of the latter two skills.

The following experimental hypotheses are offered for an initial examination of the data.

Hypothesis I

At given levels of educational attainment as determined by grade placement,

(1) the performance of native speakers is superior to that of students for whom English is a second language, and

(2) the performance of Chinese speakers does not differ from that of other E.S.L. students

on each task included in the present instrument.

Rationale

After examining the instrument, all participating E.S.L. teachers agreed that even the most proficient students were not likely to perform at the same level as native speaker peers. Certain syntactic structures were frequently identified by these teachers as not having an instructional emphasis in the E.S.L. programme.

The most readily accessible E.S.L. population was comprised, in large part, of Chinese speakers. Since all E.S.L. students selected for inclusion in this study were required to meet the same specified criteria, there is no evident reason why one linguistic group should be more proficient than others on any of the task variables. Any significant findings might reveal particular difficulties of syntactic comprehension for Chinese learners not experienced by a comparable multiethnic group of students. In other words, the present study is concerned to identify, where possible, qualitative as well as quantitative differences among groups of subsamples.

Hypothesis II

In the case of both native speakers and students for whom English is a second language, performance on each task included in the present instrument is augmented at increasing levels of educational attainment as determined by grade placement.

Rationale

Research conducted on native speakers cited in the literature review suggests this to be a plausible hypothesis. Because several E.S.L. students may have started to learn English after their elementary years of schooling, it is considered less probable that second language learners also exhibit an augmented pattern similar to that of the native speakers.

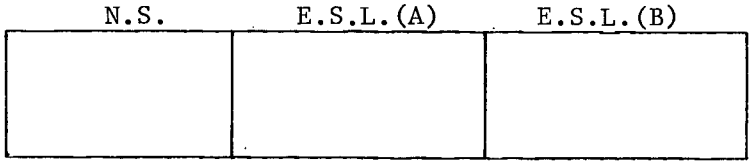
Hypothesis III

The rank order of difficulty of the sixteen subtasks of Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents) as determined by subtask mean scores does not vary among any subsamples.

Rationale

All students, regardless of their native language, are at some point in developing their comprehension of written English. It is expected that those transformational and anphoric types which are among the most difficult for those less proficient will also be relatively more difficult than other types for the more accomplished students.

Since comparison of subsamples by native language factor is of primary interest in this study, the analysis by grade interval being conducted merely to confirm a probable pattern, the remaining hypotheses are tested across three grouped samples: native speakers of English [N.S.], Chinese speakers [E.S.L.(A)], and speakers of other languages [E.S.L.(B)].



In this manner, limitations of sample size may be partially overcome while directing attention to broad contrasts.

Hypothesis IV

Combined across grade intervals,

(1) the performance of native speakers is superior to that of students for whom English is a second language, and

(2) the performance of Chinese speakers does not differ from that of other E.S.L. students

on each task and subtask included in the present instrument.

The rationale for this hypothesis is identical to that for Hypothesis I.

Hypothesis V

The rank order of difficulty of the sixteen subtasks of Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents) as determined by subtask mean scores does not vary among any subsamples combined across grade intervals.

The rationale for this hypothesis is identical to that for Hypothesis III.

Hypothesis VI

A comparison of the three measures of text comprehension exhibits a pattern of diminished performance wherein Task 4 (judgment of truth value) scores are greater than Task 3 (indication of textual locus) scores which are greater than Task 5 scores.

Rationale

If, as passage dependency studies have shown, it is possible for students to respond correctly to multiple choice items without reference to the related text at greater than chance probabilities, then it is reasonable to expect that Task 4 is easier than Task 3 which requires, for each item, that the student evaluate the propositional content of a number of sentences in the text to decide which one or two provide information sufficient to judge the truth value of the item statement.

Because Task 5 demands proficiency on both Task 3 and Task 4, it is inevitable that if Task 4 proves to be easier than Task 3 as suggested, then an individual's score for Task 5 will be lower than that attained for either of its component tasks.

Hypothesis VII

Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents) are each better predictors of Task 5 (the combined measure of locus indication and truth value judgment) than of Task 4 (judgment of truth value only.) Task 1 and Task 2 are equally good predictors of Task 5 scores.

Rationale

The model proposed to define which sequences of behaviour constitute acts of text comprehension in the present test format (q.v. Appendix "A") cites the locating of an item paraphrase in the text as crucial when short term memory fails to prompt a judgment of truth value. The sentence comparison aspect of Task 1 and the resolution nature of Task 2 are thought to be equally and more closely related to the paraphrase search embodied in Task 3 (indication of textual locus) which is a requisite for the suggested criterion of comprehension, Task 5, but not for Task 4.

Hypothesis VIII

Task 3 (indication of textual locus) is a better predictor than Task 4 (judgment of truth value) of the criterion for comprehension, Task 5.

Rationale

It will be recalled that Task 5 is a combined measure of proficiency on Task 3 and Task 4. Previous discussion of passage dependency asserts that multiple choice truth value tests (e.g., Task 4) cannot be considered as dependable measures of what for this study has been defined as an act of text comprehension. Because Task 4 is believed to have a high risk toward spurious scores, success on Task 5 is more likely to be limited by performance on Task 3.

Chapter V

PROCEDURES

Preliminary appraisal of test items

The investigator selected thirteen transformational contrasts and four anaphoric types for examination. Items constructed for each of these subtasks were reviewed by two linguists who judged their appropriateness to the category specified. Following revisions, seventy-two items were included in the final version of Test Nos. 1, 2, and 3 (i.e. Task 1 and Task 2).

Because the aforementioned tests are intended to measure syntactic understanding, vocabulary content, an established major determinant of reading comprehension scores, had to be rigorously controlled in order to assure lexical access for younger and non-native Ss. Earlier studies had concluded the use of nonsense words to afford a "purer measure" of syntactic comprehension to be an ineffective strategy owing to the interaction of syntax and semantics. Accordingly, items were constructed and reviewed to assure the propositional content of each sentence to be plausible to the Ss. Sentence topics were generally restricted to concrete objects or observable events considered to be within the realm of pupil experience and cognitive development.

Prior to the preparation of test booklets, all revised items and written directions for both the syntactic and text comprehension parts of the instrument were evaluated according to the Dale - Chall formula, essentially an index of vocabulary difficulty and sentence length. All material analyzed yielded a "corrected readability level of Grade 4.0 and below".

Format of the tests

The instrument was typewritten and photocopied to provide uniformly clear copies to all Ss who indicated their responses directly in the test booklets according to directions provided in print or by the test administrator.

To facilitate explanation of directions as well as provide Ss with suitable sample items, the instrument was divided into four sections

Test No. 1 (36 items)

This test was designed to assess recognition of two-way implication relations between paired sentences. Ss were given directions followed by two sample items:

Directions: Read both sentences.
If the two sentences mean the same thing, circle "yes".
If the two sentences DO NOT mean the same thing, circle "no".

Sample A: The boy hit the girl.
The girl was hit by the boy. (yes) no

Sample B: The boy looked at the big dog.
The big dog looked at the boy. yes (no)

The thirty-six items comprise nine subtasks of four items each which are intended to determine the S's facility with the following transformational types:

A. Passivization
B. Participle modifiers
C. Wh-- fronting
D. Relativization contrasted with clausal conjunction
E. Relativization by pronoun deletion
F. Double transformation (Relativization + Passivization)
G. Ask (query) contrasted with Tell
H. Easy to see
J. Promise contrasted with Tell

Items in Test No. 1 were arranged to correspond with the above subtasks in a rotating sequence, i.e., Item No. 1 belongs to Subtask A, Item No. 2 to Subtask B, Item No. 3 to Subtask C, etc.

Test No. 2 (16 items)

Because certain transformations of interest to the investigator could not be presented for consideration as possible two-way implication relations, an alternative format was devised to examine the S's facility with the more restrictive one-way implication relation. Again, directions and two sample items were provided:

Directions: Read the first sentence carefully.
Then read the second sentence and decide
if it is true or false.

Sample A: if: Ann and Helen walk to school together.
does it mean: Ann and Helen walk to school at the same time. (yes) no

Sample B: if: Mother said, "You must come home early."
does it mean: Mother must come home early. yes (no)

The sixteen items comprise four subtasks of four items each dealing with the following transformations:

K. Indirect speech
L. Ask (request) contrasted with Tell
M. Pseudoimperatives
N. Agentless passivization

Items in Test No. 2 were arranged consecutively in clusters, i.e., the first four items comprise Subtask K, the second four items, Subtask L, etc., in the belief that this format would minimize the cognitive shifting expected to occur when one must interpret first one transformational type, then others, only to return to another item of the first type. The

investigator was concerned to reduce the time required for response so as to maximize the opportunity for all Ss to complete the entire set of tasks in a reasonable amount of time.

Since the investigator had no reason to hypothesize that recognition of one-way implication relations is a linguistic skill distinct in its distribution from that of recognition of two-way implication relations, both Test No. 1 and Test No. 2 are combined to measure a single generalized ability, recognition of grammatical implication. This is referred to elsewhere in this report as Task 1.

Test No. 3 (20 items)

This test was designed to assess the S's capacity to identify anaphoric referents. Types chosen for the present instrument were cited in earlier discussion of anaphora. For each item, the relevant pro-form is underlined in the context. It is then repeated followed by a line provided for the S's constructed response. The task was considered to be self-explanatory through the one example offered.

Sample: I had an apple for lunch. It was good.
It apple

It was decided that any additional direction could be better provided orally by the test administrator than by written directions or further samples.

Items related to each of four subtasks were placed randomly throughout the test. Test No. 3 is referred to elsewhere in this report as Task 2.

Test No. 4 (23 items)

This test comprises two measures of text comprehension separately and in combination. All items are contained on a single page which is accompanied by another single page containing the reading passage. The latter page detaches from the test booklet so as to afford Ss continuous easy reference while responding to items. The passage consists of twelve sentences which are printed successively, one sentence to a line, regardless of sentence length. Each sentence is preceded by its number for identification as required by the following directions:

Directions: Read the story on the short paper first.
Read each question sentence below.

In the parentheses (___) write the number of the sentence, or sentences, in the story that tells you the answer.

If the question sentence is true, circle "T".

If the question sentence is false, circle "F".

If none of the sentences in the story tell you the answer to the question, put ✕ in the parentheses (___) and circle "?".

Sample A: The cave men did not use the skins of animals. (5) T (F) ?

Sample B: They lived in caves on the sides of hills where they could keep dry and warm. (2+3) (T) F ?

Sample C: The cave men built fires in front of their homes. (✕) T F (?)

The indication of textual locus portion of Test No. 4 defines Task 3. The judgment of truth value portion defines Task 4. The combination of these two components generates that measure referred to elsewhere in this report as Task 5.

Description of the sample populations

Comparison of performance among the sample populations is motivated primarily by the question: At the conclusion of an intensive programme in English language training, how similar are E.S.L. students to their native speaker peers with respect to the linguistic variables under investigation?

Because the E.S.L. populations to which the investigator had access were dominated by Chinese speaking students, it was decided to separate this group at each grade interval to determine if the performance of Chinese speakers differs, either quantitatively or qualitatively, from that of other E.S.L. students representative of a number of native languages and who had experienced similar instruction since all classes participating in the study included both Chinese and non-Chinese speakers.

It was considered appropriate to establish the following criteria for admission into the respective samples.

For native speakers:

1. Acquisition of English prior to any other language.[1]
2. No report of marked difficulty in growth in reading comprehension or of any other observed learning disability.

For E.S.L. students:

1. Completion of a minimum of 800 hours of English language training in the Vancouver schools. This was a firm requirement, apart from

---

1. This criterion eliminated the inclusion of several students enrolled in regular classes who immigrated to Canada several years ago and now claim to be more proficient in English than in their original language.

any prior study of English the student may have undertaken in his own country.

2. Assessment by the E.S.L. teacher that the student had made at least average progress in the programme, given the duration of his attendance.
3. Recommendation by the E.S.L. teacher that the student be placed in a regular class with English speaking peers on the reopening of schools the following September.[2] (The instrument was administered at the end of June.)

All Ss were enrolled in public schools under the jurisdiction of the Vancouver School Board from whom approval was granted to conduct this study. Following is a description of each subsample.

1. Native speakers -- Elementary:

   12 Ss, 6 males and 6 females, ranging in age from 9 yr 1 mo to 10 yr 5 mo, destined to enter a Grade 5 class on reopening of school. School "A" is situated in what the investigator would describe as a middle class neighbourhood with an ethnically diverse population.

2. Native speakers -- Junior Secondary

   10 Ss, 4 males and 6 females, ranging in age from 12 yr 10 mo to 14 yr 2mo, destined to enter a Grade 8 class on reopening of school by transferring

---

2. In the case of senior secondary E.S.L. students, the Ss were already enrolled in some regular courses with native speaker peers. The sample was drawn from a group of students who were required to include a special "Transitional English" course in their programmes.

from School "A" to the nearest secondary school.[3]

3. Native speakers -- Senior Secondary

   21 Ss, 9 males and 12 females, ranging in age from 14 yr 3 mo to 15 yr 10 mo, destinded to enroll in "Grade 10 English" on reopening of school. School "B" is located in an upper middle class neighbourhood which includes a small minority of immigrants.

4. E.S.L. students -- Elementary

   16 Ss, 9 males and 7 females, ranging in age from 10 yr 0 mo to 12 yr 5 mo, destined for placement in regular Grade 4, Grade 5, or Grade 6 classes on reopening of school. 10 Ss were identified to be Chinese speakers. The remaining 6 Ss include one speaker each of French, Portugese, Polish, Hungarian, Yugoslavian, and Korean.

Students in this sample were drawn from three E.S.L. reception classes at School "C" and one such class at School "D". In a reception class, usually limited to an enrollment of fifteen, the student spends most of his instructional time in a self-contained classroom under the direction of one teacher and, in some instances, a trained "teaching assistant". Arrangements vary among classes but, generally, a student does not spend more than 20% of his instructional time in other settings. Two of the three participating classes at School "C" were taught by teachers who, for the final three months of the school year, had adopted a team approach enabling them to group students according to English language proficiency. One class was designated "Intermediate", the other

---

3. In British Columbia, what is customarily the first year of junior secondary, Grade 7, is conducted as the final year of the elementary school programme.

"Advanced". The third participating class at School "C" may be characterized as multilevel. The E.S.L. programme at School "D" was divided into four phases, each under the direction of a different teacher, all of whom jointly developed a curriculum intended to maximize instructional continuity. The class participating in this study was designated as the most advanced level. Both School "C" and School "D" are located in lower - lower middle class neighbourhoods. However, one must bear in mind that the majority of students in the E.S.L. reception classes resided outside the immediate neighbourhood of the school which they attended.

5. E.S.L. students -- Junior Secondary

30 Ss, 18 males and 12 females, ranging in age from 11 yr 7 mo to 15 yr 7 mo, destined for placement in regular Grade 7, Grade 8, or Grade 9 classes on reopening of school. 23 Ss were identified to be Chinese speakers. The remaining 7 Ss include one speaker each of French, Spanish, Italian, Vietnamese, Hindi, and two Korean speakers.

In addition to the E.S.L. classes described above in 4., two other multilevel reception classes at School "C" provided Ss for this sample.

6. E.S.L. students -- Senior secondary

18 Ss, 10 males and 8 females, ranging in age from 14 yr 10 mo to 18 yr 6 mo. 12 Ss were identified to be Chinese speakers. The remaining 6 Ss include two speakers of Portugese, one speaker of Panjabi, and three Tagalog speakers.

These students were enrolled in a supplementary "Transitional English" course in addition to other courses in the regular secondary programme. On reopening of school, students in this sample would be

classified as either Grade 10 or Grade 11. School "E" is situated in a lower middle class, ethnically diverse, area of the city and, in most cases, is the facility closest to the student's residence.

It will be noted that the number of Ss in each subsample is unequal. While the investigator would have preferred to increase the sample size of native speakers, this was not feasible as it would have necessitated accessing a number of classes at a busy time of the school year. It is also considered that the linguistic variables examined by the instruments are aspects of language competence which, for a native speaker population of a given school grade interval, would not be greatly influenced by curriculum content or methods of teaching experienced by the student during the school year. On the other hand, considerable effort was expended to procure a sizable E.S.L. sample from several different sources within the school system so as to randomize the effects of strengths and weaknesses of specific programmes since one might reasonably propose that the methodology and curriculum selected by the E.S.L. teacher throughout the school year could contribute substantially to the variance in students' performance on measures of the particular linguistic variables under investigation.

Administration of the instrument

In all but one case the tests were administered by the investigator to Ss in the elementary and junior secondary samples in their usual classroom setting. The one exception, an elementary / junior secondary E.S.L. class from which eight pupils were deemed to be suitable Ss for

this study, and the senior secondary samples completed the tests under the supervision of their regular English teacher who had been briefed by the investigator on the purpose of the study, what each test was designed to measure, and specific problems which might arise in administration.

All administrations were conducted in June, 1977, during the final three weeks of the school year. Scheduling restraints and the number of participating classes in dispersed locations necessitated completion of all tasks in one sitting. This condition may have been a contributing factor to the poor performance of some Ss. However, 90.7% of the native speakers and 96.9% of the E.S.L. students who met the criteria for admission into the sample responded to all, or nearly all, items on the final tasks.

Pilot testing of the instrument on a small sample of elementary students (not participants in this study) who were known to vary widely in reading comprehension skills indicated the probability that most Ss would be able to complete all tasks in less than fifty minutes. Accordingly, all Ss were advised that their performance would not be timed and that they would be granted adequate time to complete all tasks carefully. Very few Ss required more than fifty minutes for completion.

Owing to individual differences in rate of response to various types of items, it was decided to present directions and review examples of all tasks before allowing Ss to commence the first test. Supplementary oral directions appropriate to each of the four tests in the instrument were as follows:

Tests No. 1, No. 2, and No. 3

All classes with the exception of the senior secondary native speakers were pre-tested for recognition of gender of the common personal

names used in test items. E.S.L. classes were posted a list of such names.

Test No. 3

Ss were cautioned that despite a single short line on which to write their responses, the most appropriate referent for the underlined pro-form might possibly comprise several words. An example, additonal to that presented in the test booklet, was presented:

> Jack will not come to school today. He said so yesterday.
>
> so = that Jack will not come to school

Scoring of the tests

Each test booklet was hand coded to prepare input data for a LERTAP 2.0 computer programme (Nelson, 1974) to compute and tabulate raw scores, present descriptive statistics, and furnish item analyses. Coding of booklets was double checked and a random sampling of booklets representing 20% of the entire sample revealed no clerical errors when verified by an external source. All keypunching was later verified either visually or mechanically.

In the case of Tests No. 1 (Task 1), No. 2 (Task 1), and the Truth Value portion of No. 4 (Task 4), answer keys were prepared in advance of scoring.

Test No. 3 (Task 2) in which Ss were required to identify anaphoric referents is characterized by a "constructed response" format. The investigator evaluated all responses for this test in two concentrated sessions approximately two days apart in order to establish a consistent standard in setting minimum criteria for acceptable responses. As no key could be set, responses were judged somewhat subjectively for evidence of grasp of the essential idea rather than for precision in articulating an utterance that

could be directly substituted into the pro-form position. By way of example, consider Item No. 5:

> The boys played ball very hard. This is what won the game.
>
> This _____

Typical responses judged acceptable:

> playing ball very hard
>
> playing very hard
>
> playing hard
>
> played hard

Typical responses judged unacceptable:

> playing ball
>
> very hard

An inspection of responses to the textual locus portion of Test No. 4 (Task 3) indicated the need to depart from a strict adherence to the predetermined key as the investigator concluded the truth value of the item statement could in certain cases be derived from a combination of sentences other than what was suggested by the key. The following example is illustrative of the problems incurred in constructing and keying a test of text comprehension for which there is no prototype in the literature.

Item No. 4 The picture stories about cave men tell us new things. (___) T F ?

Keyed response: (12) They tell us things about cave men that we never knew before.

Alternate responses: (10) The picture stories are still there.
(12) They tell us things about cave men that we never knew before.

(8) They drew their pictures on the stone walls of caves.
(12) They tell us things about cave men that we never knew before.

After scoring this test, the investigator rechecked all test booklets to verify that a consistent standard for evaluating responses was maintained for all subsamples.

Chapter VI

RESULTS AND DISCUSSION

Statistical properties of the instrument

This preliminary discussion is intended to outline those steps necessary to assess the validity of the instrument in its present form.

Task 1

Task 1 (recognition of grammatical implication) comprises nine subtasks pertaining to two-way implication relations (Test No. 1) and four subtasks pertaining to one-way implication relations (Test No. 2). Each subtask contains four items for a total of 52 items.

The preliminary output of the LERTAP 2.0 programme indicated that one subtask, N: Agentless passivization, has a low correlation of .015 with Task 1 as a whole thereby indicating that the subtask makes no substantive contribution to the variable under investigation. Further evidence of the inappropriateness of Subtask N lies in the observation that two items bear negative point biserial correlations with total task scores. This means that, as a group, those Ss who answered the items correctly earned lower scores on Task 1 than those who answered incorrectly. This particular subtask yielded a mean score of 2.04 for the entire sample, noticeably lower than that of any other subtask associated with Task 1. Moreover, the highest scores were earned by E.S.L. students at the lower grade intervals. Consequently, the decision was made to remove Subtask N from any further analysis of the instrument. The items are included in Appendix "B" and the subtask is discussed in a later section of this report.

The remainder of Task 1 contains twelve subtasks, a total of 48 items. The mean task score obtained by the entire sample population is 40.04, approximately 83.4% of the maximum possible score. The standard deviation is 5.17 and the range is 25 to 48, thus indicating a negative skewness characteristic of a measure of mastery learning as might be expected of any aspect of linguistic competence.

Since the validity of any test is limited by its reliability, it is important to ascertain that measures of reliability are reasonably high. This is to say that a researcher wants to be certain that each item on his test contributes to a measure of the same variable. The LERTAP 2.0 programme provides a number of statistics to monitor internal reliability. Four have been selected for use in the analysis of data.

The point biserial correlation of each item with the subtask score (PB-ST) offers some indication of the extent to which a particular item acts in concert with the other items of the subtask to measure what is conceived to be a specific variable. PB-ST correlations among the various subtasks of Task 1 range from .20 to .82 with a median PB-ST of .60.

The point biserial correlation of each item with the total task score (PB-TT) offers some indication of the extent to which a particular item acts in concert with the other items appearing on the total task to measure what is usually conceived to be a more generalized variable that logically subsumes the more particularized variables upon which the separate subtasks are constructed. If this is in fact the case, the PB-TT correlations will encompass a somewhat lower range than their associated PB-ST coefficients. The 48 PB-TT coefficients for Task 1 range from −.05 to .57 with a median PB-TT of .28.

The Hoyt estimate of reliability is designed to measure the internal consistency of a test and is comparable to the also popular Kuder - Richardson KR 20 statistic. Nelson (1974:260) explains:

> Internal consistency is an estimate of the extent to which each test item taps whatever the test is measuring. We might consider each test item as a sample test from the total domain; then the internal consistency is roughly equivalent to the average correlation between all pairs of items (or sample tests).

Hoyt estimates for the subtasks of Task 1 are generally low, ranging from .00 to .68 with a median reliability coefficient of .31. This is to be expected in view of the fact that each subtask contains only four items. The Hoyt estimate for the total task is a more dependable indicator as it is based upon a total of 48 items. For Task 1, this coefficient is .80.

Cronbach's alpha provides a more rigorous verification of the proposition that the subtasks each contribute to the measurement of a single generalized variable.

> Coefficient alpha is an index of the consistency of the subtests, or, the degree to which the subtests tend to measure the same thing. Cronbach suggests that this alpha is an index of how much the total test score reflects "common elements rather than a hodgepodge of elements each specific to one subtest." (Nelson, 1974:280).

The Cronbach's alpha for the composite of the subtasks of Task 1 is .75.

Reviewing the four aforementioned indices in accord with generally accepted standards for test construction and in specific comparison with statistics reported for the previously cited instruments purporting to test recognition of semantic equivalence of syntactic structures, it is reasonable to conclude that the statistical properties of the tests for Task 1 are adequate.

One of the most useful features of the LERTAP 2.0 programme is the correlation matrix assembled from scores on all subtasks, the total task, and an external criterion. Inter-subtask correlations for Task 1 range from .033 to .370. The correlations between each of the subtasks and the total task vary from a low of .340 to a high of .713; only one coefficient lies in the range of the inter-subtask correlations. From this result it is conceivable that the twelve subtasks of Task 1 each measure a unique aspect of linguistic competence, all of which combine to contribute to a generalized language skill, recognition of grammatical implication.

A correlation coefficient is computed between each subtask score and an external criterion; in this case, the variable is age recorded in months. The correlations related to Task 1 are all very moderate, ranging from ⁻.009 to .280. The correlation between age and total task score is .278. These results support the contention that while Ss' performance may appear to improve at higher grade intervals and, therefore, with increasing age, there is still a wide variation in syntactic comprehension among a group of Ss of a given age.

One final point of interest is the matter of possible sex differences in performance. A t-test of independent means produced the following values:

	N	Mean	st dev	t value	prob
Males	56	39.32	5.14		
Females	51	40.82	5.14	⁻1.51	not sig.

It is therefore concluded that, for the entire sample treated as a single group, sex is not a significant factor in performance on Task 1.

Task 2

Task 2 (identification of anaphoric referents) comprises four subtasks (Test No. 3) and contains a total of 20 items. The mean task score obtained by the entire sample population is 14.10, approximately 70.5% of the maximum possible score. The standard deviation is 5.27 and the range is 0 to 20. As with Task 1, the distribution is negatively skewed.

Point biserial correlations among the subtasks of Task 2 range from .43 to .84 with a median PB-ST of .74.

Point biserial correlations relating each item to the total task range from .34 to .80 with a median PB-TT of .62.

Hoyt estimates of reliability for the subtasks of Task 2 range from .63 to .79. These tend to be much higher than the estimates for the subtasks of Task 1. The Hoyt estimate for the total task based on 20 items is .91. The Cronbach's alpha for the composite of the subtasks of Task 2 is .87.

It, therefore, appears that the statistical properties of Task 2 are adequate and superior to those of Task 1.

Inter-subtask correlations for Task 2 range from .439 to .794. The correlations between each of the subtasks and the total task vary from a low of .674 to a high of .930. As with Task 1, only one coefficient lies in the range of the inter-subtask correlations. Hence, it again seems conceivable that each subtask measures a somewhat unique aspect of linguistic competence which when combined with the other subtasks contributes to a more generalized language skill, the ability to correctly identify anaphoric referents.

The correlation between age in months and scores on each of the subtasks of Task 2 are all moderate, ranging from .276 to .381. The correlation between age and total task score is .369. These coefficients are somewhat higher than those reported for Task 1. Nevertheless, the suggestion of wide variation in linguistic competence at a given age is supported.

Finally, with regard to the possibility of sex differences in performance, a t-test of independent means produced the following values:

	N	Mean	st dev	t value	prob
Males	56	12.95	5.58		
Females	51	15.37	4.64	-2.43	$<.025$

It is therefore concluded that, for the entire sample treated as a single group, sex is a significant factor favouring the performance of female Ss on Task 2.

Task 3

Task 3 contains 23 items and constitutes the indication of textual locus component of Test No. 4. The mean task score obtained by the entire sample population is 15.96, approximately 69.4% of the maximum possible score. The standard deviation is 3.50 and the range is 2 to 21. This distribution of scores, unlike those for Task 1 and Task 2, more closely appraoches normality.

The point biserial correlations for the 23 items of Task 3 range from .08 to .60 with a median PB-TT of .39. The Hoyt estimate of reliability for Task 3 is .72. The statistical properties of Task 3, while not as strong as those for Tasks 1 and 2, may nevertheless be considered adequate.

The correlation between Task 3 scores and the external criterion

of age is .466, a moderate relationship.

In the matter of possible sex differences in performance, a t-test of independent means produced the following values:

	N	Mean	st dev	t value	prob
Males	53	15.98	3.19		
Females	48	15.94	3.85	0.06	not sig.

It is therefore concluded that, for the entire sample treated as a single group, sex is not a significant factor in performance on Task 3.

Task 4

Task 4 contains 23 items and constitutes the judgment of truth value component of Test No. 4. The mean task score obtained by the entire sample population is 16.42, approximately 71.4% of the maximum possible score. The standard deviation is 3.27 and the range is 7 to 22. This distribution of scores, unlike those for Task 1 and Task 2, more closely approaches normality.

The point biserial correlations for the 23 items of Task 4 range from $^{-}$.05 to .60 with a median PB-TT of .39. The Hoyt estimate of reliability for Task 4 is .66. The statistical properties of Task 4, while not as strong as those for Tasks 1 and 2, may nevertheless be considered adequate.

The correlation between Task 4 scores and the external criterion of age is .304, a very moderate relationship.

In the matter of possible sex differences in performance, a t-test of independent means produced the following values:

	N	Mean	st dev	t value	prob
Males	56	16.45	2.83		
Females	50	16.38	3.73	0.10	not sig.

It is therefore concluded that, for the entire sample treated as a single group, sex is not a significant factor in performance on Task 4.

Task 5

Task 5 contains 23 items and is based on the combination of textual locus indication (Task 3) and truth value judgment (Task 4). This combination of both tasks has been designated as the criterion measure of text comprehension. The mean task score obtained by the entire sample population is 14.58, approximately 63.0% of the maximum possible score. The standard deviation is 3.65 and the range is 1 to 21. This distribution of scores is similar to those for Task 3 and Task 4 and, unlike the distributions for Task 1 and Task 2, more closely approaches normality.

The point biserial correlations for the 23 items of Task 5 range from .08 to .58 with a median PB-TT of .39. The Hoyt estimate of reliability for Task 5 is .72. The statistical properties of Task 5, while not as strong as those for Tasks 1 and 2, may nevertheless be considered adequate.

The correlation between Task 5 scores and the external criterion of age is .414, a moderate relationship.

In the matter of possible sex differences in performance, a t-test of independent means produced the following values:

	N	Mean	st dev	t value	prob
Males	53	14.47	3.23		
Females	48	14.52	4.07	-0.07	not sig.

It is therefore concluded that, for the entire sample treated as a single group, sex is not a significant factor in performance on Task 5.

Having completed a preliminary survey of the general statistical properties of the research instrument utilizing the entire sample population some attention must next be focused upon these same properties as they relate to specific subsamples within the larger sample. Given the limitations of sample size, it was considered appropriate to examine only two subsamples: all native speakers and all E.S.L. students. An exception is the matter of accurately reporting estimates of reliability. To present a single coefficient for each of these two subsamples, it was considered necessary to equate the number of Ss at each grade interval so as not to bias the Hoyt coefficient by allowing any one grade interval to be over-represented in the calculation. This reduction in sample size was accomplished through elimination of Ss based on a table of random numbers.

More accurate is the reporting of separate internal reliability estimates for each of six sample populations. In this way, the comparative suitability of different tasks for specified populations can be readily ascertained. This should be noted as an important precaution against misapplication of a test upon an unsuited population. Coefficients reported herein are, however, tentative. Internal reliability estimates obtained for

certain subsamples may be improved when applied to a larger sample size. For example, the Hoyt coefficient for Native Speakers, Task 3 is .81 when all 39 Ss completing the task are included in the sample. By reducing the number of Ss to 21 to maintain equal distribution across grade intervals, the estimate of reliability drops to .47.

Two points of interest are to be observed from the comparative data: (1) the only significant sex difference in performance is confined to that of E.S.L. students on Task 2, and (2) age - task correlations for native speakers are consistently higher than for E.S.L. students in accord with the suggestion made elsewhere that more linear developmental patterns should be apparent for native speakers. The limited magnitude of the correlations, however, tends to verify that considerable variation in performance is to be encountered at any given age.

Relevant indices as outlined are set out in Tables 1 - 5 to facilitate direct comparisons between the native and non-native samples.

Problems in statistical analysis of the data

Having evaluated the statistical properties of each task in relation to the entire sample population, the investigator wished to make certain comparisons among subsamples. It was originally intended to apply a standard two-way analysis of variance to the data as a first step to locating any significant differences. One mathematical assumption upon which the analysis of variance procedure is based is that of homogeneity of sample variances. Tests of homogeneity were therefore applied to the data at hand: maximum variance / minimum variance ratio and Bartlett - Box F, the latter generally considered to be the most appropriate of all

	N.S.	E.S.L.
Mean	42.86	38.23
St dev.	4.52	4.73
Range of scores	25 to 48	28 to 48
PB - ST range	.00 to .91	.00 to .84
Median PB - ST	.64	.57
PB - TT range	-.03 to .74	-.01 to .56
Median PB - TT	.35	.27
Hoyt reliability*	.69	.73
Cronbach's alpha*	.63	.69
Inter-subtask corrs	-.207 to . 608	-.053 to .417
Subtask - Total task corrs	.293 to .734	.370 to .707
Age - Total task corr	.507	.302

* based on equal number of Ss at each Grade interval

Hoyt estimate of reliability (Cronbach's alpha)
for six sample populations**

	Elem	JrSec	SrSec
N.S.	.86 (.82)	.64 (.58)	.32 (.21)
E.S.L.	.79 (.79)	.64 (.55)	.66 (.64)

** based on all Ss completing the task

---

t-test for sex differences

		N	Mean	st dev	t value	prob.
N.S.	Males	19	43.16	3.61		
	Females	24	42.63	5.19	0.38	not sig.
E.S.L.	Males	37	37.35	4.69		
	Females	27	39.22	4.61	-1.59	not sig.

Table 1: Test statistics for Task 1: Native Speakers and E.S.L. Students compared

	N.S.	E.S.L.
Mean	14.72	13.82
St dev	6.53	4.26
Range of scores	0 to 20	1 to 20
PB - ST range	.53 to .93	.38 to .80
Median PB - ST	.85	.70
PB - TT range	.49 to .92	.25 to .77
Median PB - TT	.81	.51
Hoyt reliability*	.96	.84
Cronbach's alpha*	.92	.80
Inter-subtask corrs	.540 to .916	.318 to .639
Subtask - Total task corrs	.720 to .970	.655 to .869
Age - Total task corr	.685	.132

* based on equal number of Ss at each Grade interval

Hoyt estimate of reliability (Cronbach's alpha)
for six sample populations**

	Elem	JrSec	SrSec
N.S.	.97 (.94)	.91 (.85)	.10 (-.37)
E.S.L.	.88 (.83)	.72 (.71)	.88 (.82)

** based on all Ss completing the task

---

t-test for sex differences

		N	Mean	st dev	t value	prob.
N.S.	Males	19	13.42	6.83		
	Females	24	15.54	6.23	-1.06	not sig.
E.S.L.	Males	37	12.70	4.91		
	Females	27	15.22	2.65	-2.64	$<.025$

Table 2: Test statistics for Task 2: Native Speakers and E.S.L. Students compared

	N.S.	E.S.L.
Mean	15.56	16.29
St dev	4.26	3.03
Range of scores	2 to 21	10 to 21
PB - Task range	-.02 to .82	.02 to .63
Median PB - Task	.47	.35
Hoyt reliability*	.47	.62
Age - Task corr	.596	.365

* based on equal number of Ss at each Grade interval

Hoyt estimate of reliability for six sample populations**

	Elem	JrSec	SrSec
N.S.	.87	.29	.16
E.S.L.	.23	.68	.67

** based on all Ss completing the task

---

t-test for sex differences

		N	Mean	st dev	t value	prob
N.S.	Males	18	15.61	4.05		
	Females	21	15.52	4.54	0.06	not sig.
E.S.L.	Males	35	16.17	2.70		
	Females	27	16.26	3.28	-0.12	not sig.

Table 3: Test statistics for Task 3: Native Speakers and E.S.L. Students compared

	N.S.	E.S.L.
Mean	17.26	16.22
St dev	3.27	3.13
Range of scores	7 to 22	8 to 21
PB - Task range	-.09 to .71	-.08 to .61
Median PB - Task	.40	.35
Hoyt reliability*	.70	.57
Age - Task corr	.470	.310

* based on equal number of Ss at each Grade interval

Hoyt estimate of reliability for six sample populations**

	Elem	JrSec	SrSec
N.S.	.73	.70	.46
E.S.L.	.43	.60	.75

** based on all Ss completing the task

---

t-test for sex differences

		N	Mean	st dev	t value	prob
N.S.	Males	19	17,84	3.29		
	Females	23	16.26	3.62	1.47	not sig.
E.S.L.	Males	37	15.73	2.29		
	Females	27	16.48	3.89	-0.90	not sig.

Table 4: Test statistics for Task 4: Native Speakers and E.S.L. Students compared

	N.S.	E.S.L.
Mean	14.62	14.54
St dev	4.34	3.25
Range of scores	1 to 21	8 to 20
PB - Task range	.15 to .78	.01 to .63
Median PB - Task	.47	.35
Hoyt reliability	.46	.80
Age - Task corr	.602	.288

* based on equal number of Ss at each Grade interval

Hoyt estimate of reliability for six sample populations**

	Elem	JrSec	SrSec
N.S.	.85	.12	.42
E.S.L.	.45	.64	.77

** based on all Ss completing the task

---

t-test for sex differences

		N	Mean	st dev	t value	prob
N.S.	Males	18	14.94	4.43		
	Females	21	14.29	4.41	0.47	not sig.
E.S.L.	Males	35	14.23	2.53		
	Females	27	14.70	3.87	-0.58	not sig.

Table 5: Test statistics for Task 5: Native Speakers and E.S.L. Students compared

such tests in the case of unequal cells. (Winer, 1962:95). As can be noted in Tables 6 - 10, the homogeneity of variances requirement is clearly violated with respect to all tasks on which the subsamples are to be compared. Ferguson (1976:234) warns: "Gross departure from homogeneity may lead to results which are seriously in error." and advises that, under certain conditions, mathematical transformations of the data may be applied to reduce the variation among variances. A search for a suitable transformation, however, did not prove fruitful in this instance.

Consideration was given to adding or deleting Ss to effect uniformity of sample size. This course was rejected for two reasons, one statistical, the other empirical. First, a proper selection of additional or retained Ss would most probably resemble the present subsamples and thereby fail to reduce the range of variances. While the condition of equal sample size does generally abate the demand for homogeneity, a ratio as great as 20:1 among variances cannot be considered to lie within the tolerance of the analysis of variance procedure. Secondly, it has been well established that, for native speakers of English, acquisition of syntactic structures, particularly some of those selected for this study, occurs over a broad age range (Chomsky, 1969). An inspection of Tables 6 and 7 relating to performance on Tasks 1 and 2, respectively, will reveal a pattern of decreasing variances across grade intervals (grade placement being highly correlated with age) for native speakers. This phenomenon is consistent with earlier reports which suggest that one might expect to encounter greater variations in linguistic abilities among younger pupils. Children learning English as a second language might be bound by similar constraints in their ability to grasp the underlying meaning of

certain syntactic forms which may, or may not, be presented in the programme of second language instruction. Like their native speaker counterparts, the E.S.L. subsamples exhibit a pattern of decreasing variance in the case of Task 1 (Table 6). This trend, however, is not evident in a comparison among the E.S.L. subsamples with respect to Task 2 (Table 7).

It therefore seems reasonable to postulate that, in some instances, a lack of homogeneity of variances is an inevitable occurance when one chooses to make comparisons among widely disparate populations with respect to any variable that may be characterized as an aspect of linguistic competence.

In view of these problematic results, the investigator elected to proceed with an analysis of the data by use of appropriate non-parametric tests.

Comparison of samples by Language group and Grade interval

The following sequence is used in presenting summary comparisons of the performance of all subsamples:

(1) Mean scores, variances, and number of Ss, listed in this format:

39.17 ← Subsample mean score

12 39.42 ← Subsample variance in scores

↑ Number of Ss in the subsample

(2) Indices of homogeneity of variances

(3) Mean ranks: Kruskal - Wallis non-parametric one-way analysis of variance, corrected for ties

(4) Summary of significant differences calculated according to a procedure developed by Dunn (1964) for the purpose of simultaneously conducting multiple comparisons from a single set of ranked data.

(1)

	Elem		JrSec		SrSec	
N.S.	39.17		43.20		44.81	
	12	39.42	10	10.84	21	4.06
E.S.L.(A)	37.50		36.74		40.17	
	10	19.83	23	20.29	12	13.97
E.S.L.(B)	37.83		36.86		42.33	
	6	49.37	7	13.48	6	15.47

(2) Maximum s^2 / Minimum s^2 = 12.154

Bartlett - Box F = 2.907, $p < .01$

(3)

	Elem	JrSec	SrSec
N.S.	50.38	73.10	85.12
E.S.L.(A)	37.10	32.54	52.58
E.S.L.(B)	45.17	32.43	67.75

Kruskal-Wallis ANOVA, $\chi^2 = 44.323$, $p < .001$

(4)

JrSec	N.S.	v.	JrSec E.S.L.(A)	$p < .001$
JrSec	N.S.	v.	JrSec E.S.L.(B)	$p < .01$
SrSec	N.S.	v.	SrSec E.S.L.(A)	$p < .01$
Elem	N.S.	v.	JrSec N.S.	$p < .05$
Elem	N.S.	v.	SrSec N.S.	$p < .001$
JrSec	E.S.L.(A)	v.	SrSec E.S.L.(A)	$p < .05$
JrSec	E.S.L.(B)	v.	SrSec E.S.L.(B)	$p < .025$

Table 6: Comparison of performance on Task 1
by Language group and Grade interval

(1)

	Elem		JrSec		SrSec	
N.S.	7.50		15.70		18.14	
	12	57.18	10	23.57	21	2.93
E.S.L.(A)	13.80		15.22		14.33	
	10	12.40	23	4.72	12	28.61
E.S.L.(B)	9.00		13.29		12.33	
	6	34.80	7	26.24	6	17.07

(2) Maximum s^2 / Minimum s^2 = 19.525

Bartlett - Box F = 5.797, $p < .001$

(3)

	Elem	JrSec	SrSec
N.S.	28.21	66.10	84.60
E.S.L.(A)	43.90	52.67	56.46
E.S.L.(B)	24.33	45.36	35.08

Kruskal-Wallis ANOVA, χ^2 = 40.035, $p < .001$

(4)

SrSec N.S.	v.	SrSec E.S.L.(A)	$p < .01$
SrSec N.S.	v.	SrSec E.S.L.(B)	$p < .001$
Elem N.S.	v.	JrSec N.S.	$p < .01$
Elem N.S.	v.	SrSec N.S.	$p < .001$

Table 7: Comparison of performance on Task 2
by Language group and Grade interval

(1)

	Elem		JrSec		SrSec	
N.S.	12.00		14.71		18.00	
	12	28.55	7	5.24	20	3.58
E.S.L.(A)	15.40		16.30		17.45	
	10	2.93	23	10.31	11	3.87
E.S.L.(B)	14.00		15.83		17.50	
	6	6.80	6	11.77	6	15.50

(2) Maximum s^2 / Minimum s^2 = 9.731

Bartlett - Box F = 3.125, $p < .01$

(3)

	Elem	JrSec	SrSec
N.S.	26.17	34.21	69.67
E.S.L.(A)	39.50	53.50	65.73
E.S.L.(B)	29.75	46.83	66.00

Kruskal-Wallis ANOVA, χ^2 = 28.735, $p < .001$

(4)

Elem	N.S.	v.	SrSec N.S.	$p < .001$
JrSec	N.S.	v.	SrSec N.S.	$p < .001$
Elem	E.S.L.(A)	v.	SrSec E.S.L.(A)	$p < .025$
Elem	E.S.L.(B)	v.	SrSec E.S.L.(B)	$p < .025$

Table 8: Comparison of performance on Task 3
by Language group and Grade interval

(1)

	Elem		JrSec		SrSec	
N.S.	15.00		16.00		18.65	
	12	16.73	10	13.56	20	4.77
E.S.L.(A)	15.40		16.39		16.83	
	10	8.27	23	7.98	12	8.33
E.S.L.(B)	13.83		15.71		16.83	
	6	4.17	7	12.57	6	20.97

(2) Maximum s^2 / Minimum s^2 = 5.032

Bartlett - Box F = 1.263, not sig.

(3)

	Elem	JrSec	SrSec
N.S.	42.25	50.20	74.25
E.S.L.(A)	42.45	52.59	57.13
E.S.L.(B)	25.83	46.50	62.83

Kruskal-Wallis ANOVA, χ^2 = 18.323, $p < .025$

(4)

Elem N.S.	v.	SrSec N.S.	$p < .01$
JrSec N.S.	v.	SrSec N.S.	$p < .025$
Elem E.S.L.(B)	v.	SrSec E.S.L.(B)	$p < .025$

Table 9: Comparison of performance on Task 4
by Language group and Grade interval

(1)

	Elem		JrSec		SrSec	
N.S.	10.83		14.43		16.90	
	12	27.42	7	4.29	20	5.67
E.S.L.(A)	14.10		14.57		15.09	
	10	6.10	23	9.44	11	7.89
E.S.L.(B)	12.17		14.17		15.50	
	6	7.37	6	15.37	6	24.30

(2) Maximum s^2 / Minimum s^2 = 6.399

Bartlett - Box F = 1.975, $p < .05$

(3)

	Elem	JrSec	SrSec
N.S.	29.54	46.36	71.07
E.S.L.(A)	45.10	50.43	56.45
E.S.L.(B)	28.17	46.83	61.42

Kruskal-Wallis ANOVA, χ^2 = 21.538, $p < .01$

(4)

Elem N.S.	v.	SrSec N.S.	$p < .001$
JrSec N.S.	v.	SrSec N.S.	$p < .025$
Elem E.S.L.(B)	v.	SrSec E.S.L.(B)	$p < .025$

Table 10: Comparison of performance on Task 5
by Language group and Grade interval

Hypothesis I : Results

> At given levels of educational attainment as determined by grade placement,
>
> (1) the performance of native speakers is superior to that of students for whom English is a second language, and
>
> (2) the performance of Chinese speakers does not differ from that of other E.S.L. students
>
> on each task included in the present instrument.

An inspection of Task 1 (recognition of grammatical implication) results shows that at each grade interval the performance of native speakers is superior to that of either sample of E.S.L. students. Of six possible comparisons, only three, at the secondary grade intervals, are statistically significant.

Chinese speakers at each grade interval score slightly lower than other E.S.L. students but these differences are not significant.

Task 2 (identification of anaphoric referents) results indicate native speakers superior to E.S.L. students only at the secondary intervals, with the two contrasts for senior secondary subsamples statistically significant. Native speakers at the elementary interval appear to be markedly inferior in this skill although no significant difference can be established with either of the E.S.L. samples.

Chinese speakers at all grade intervals clearly score higher than other E.S.L. students but none of the differences prove significant.

On Task 3 (indication of textual locus), native speakers are superior to the E.S.L. samples only at the senior secondary interval.

Chinese speakers score somewhat higher than other E.S.L. students only at the two lower intervals.

None of these differences cited for Task 3 are statistically significant.

Task 4 (judgment of truth value) results mark only senior secondary native speakers superior to both E.S.L. samples. At the two lower intervals native speakers are superior only to the non-Chinese group of E.S.L. students.

Chinese speakers at the two lower intervals score higher than other E.S.L. students, there being a tie between the two senior secondary samples.

None of these differences cited for Task 4 are statistically significant.

Finally, the criterion Task 5 results show a pattern wherein native speakers are inferior to either E.S.L. sample at the elementary interval, rank between Chinese and the other E.S.L. students at the junior secondary interval, and score higher than either E.S.L. sample at the senior secondary interval.

Chinese speakers at the two lower intervals score over other E.S.L. students, the trend being reversed at the senior secondary interval.

None of these differences cited for Task 5 are statistically significant.

Hypothesis I : Conclusions

Regarding the tests of syntactic comprehension outside of a text, results of Task 1 and Task 2 do not fully support the hypothesis of native speaker superiority across grade intervals. The hypothesis of no difference between Chinese speakers and other E.S.L. students on measures of these skills is firmly supported.

With respect to the measures of text comprehension, results of Task 3, Task 4, and Task 5 all fail to confirm the hypothesis of native speaker superiority while clearly supporting the hypothesis of no difference between Chinese speakers and other E.S.L. students.

Hypothesis I : Discussion

Analysis of the comparative data for native speakers and second language students indicates that, after 800 hours of instruction, E.S.L. students at the secondary grade intervals are more likely to underperform their native speakers peers than E.S.L. pupils in the elementary school. For the first two tasks of syntactic comprehension, these differences are statistically significant in most cases. This observation lends some support to the contention that younger children tend to progress better in learning a second language than do older students. It may be more efficient in achieving native-like competencies to promote entry into an E.S.L. programme at an early age.

Comparisons across grade intervals reveal that even after attending an E.S.L. programme over a prolonged period of time, students do not recognize grammatical implication relations as readily as their native speaker peers. Results for the anaphoric identification task are less clear. An apparent anomaly is evident in the comparative ability of the elementary subsamples, the E.S.L. samples outperforming native speakers. Some suggestion was made earlier that young children may operate under severe memory constraints which inhibit anaphoric resolution. This assertion may be true but whether the age advantage of the elementary E.S.L. subsample over the native speaker counterpart, 11 yrs 1 mo and 10 yrs 0 mo, respectively, can account for the former's higher (though

not statistically significant) mean score on Task 2 is open to question. A more plausible explanation may lie in a conscious recognition, often reflected in E.S.L. curricula, of the need for focused instruction in anaphoric structures.

Regarding the measures of text comprehension -- Tasks 3, 4, and 5, it is noteworthy that no significant differences between native speakers and E.S.L. students can be established at any grade interval. This finding strongly implies that a high level of recognition of grammatical implication relations and the ability to identify anaphoric referents are not necessary to the attainment of native-like proficiency in text comprehension tasks. It does not, however, preclude the possibility that these linguistic skills may have a facilitating effect on such tasks.

As a final point of discussion, it is interesting to compare certain results of this study with those reported by van Metre (1974) who found for Grade 3 pupils that ability to comprehend structures examined by Chomsky (1969) does not discriminate between native (monolingual) and non-native (bilingual) pupils but rather is a better predictor of reading comprehension scores. When comparison of the present results is restricted to the elementary grade interval only, which Ss are most similar to those of van Metre, it is noted that the superior performance of native speakers on Task 1 is not statistically significant. Additional analysis of the data indicates that for secondary students, Task 1 discriminates between high and low scorers on text comprehension (Task 5) as well as between native speakers and E.S.L. students. By isolating the Chomsky based items -- Subtasks G, H, J, L, and R -- from the remainder of Tasks 1 and 2, it can be

observed that these function to discriminate between high and low comprehenders (Task 5) among the elementary subsamples without separating the E.S.L. students from the native speakers. These results are remarkably consistent with van Metre, especially when it is recalled that her sample comprised pupils at high and low extremes of measured reading comprehension in contrast to the present study whose Ss are more homogeneous in text comprehension ability as evidenced by the unimodal distribution of Task 5 scores.

Hypothesis II : Results

> In the case of both native speakers and students for whom English is a second language, performance on each task included in the present instrument is augmented at increasing levels of educational attainment as determined by grade placement.

The anticipated gradient is evident in the Task 1 (recognition of grammatical implication) mean scores of native speakers, the elementary subsample performing significantly below either of the secondary intervals. The gradient for E.S.L. samples is interrupted by depressed means for the junior secondary intervals whose performance is significantly inferior only to their senior secondary counterparts.

Results for Task 2 (identification of anaphoric refernts) are very similar to those for Task 1 in the case of native speakers -- a positive gradient with the performance of the elementary subsample significantly inferior to the secondary intervals. Results for E.S.L. students differ from the preceding task insofar as scores for both of the junior secondary intervals are elevated from a linear gradient. No significant differences are reported between any intervals of either E.S.L. sample.

Task 3 (indication of textual locus) shows unbroken gradients for all three samples. In each case, the senior secondary interval is significantly superior to the elementary; for native speakers, there is also a significant difference between the junior secondary and the senior secondary intervals.

Task 4 (judgment of truth value) gradients are defined for all three samples. Performance of senior secondary native speakers is significantly superior to that of the lower intervals. The only significant difference found across grade intervals for E.S.L. students is that between the elementary and senior secondary non-Chinese learners.

Gradients for Task 5, the comprehension criterion, also show no deviations for any of the samples grouped by native language factor. Again, significant differences are confined to native speakers, senior secondary outperforming both of the lower intervals, and a contrast between elementary and senior secondary E.S.L. speakers of languages other than Chinese.

Hypothesis II : Conclusions

The hypothesis of augmented performance across grade intervals is generally confirmed for native speakers on all tasks although not all contrasts are statistically significant. Regarding E.S.L. students, similar results obtain for the tasks of text comprehension. Discounting small differences in mean scores which are not statistically significant, like gradients may evolve for the first two tasks of syntactic comprehension as well. It will be recalled, however, that certain differences between first and second language acquisition could contribute to deviations in a gradient curve for an E.S.L. sample across grade intervals while a positive progression would be a definite expectation for native speakers.

	N. S.			E.S.L.(A)			E.S.L.(B)		
	Elem	JrSec	SrSec	Elem	JrSec	SrSec	Elem	JrSec	SrSec
A. Passivization	3.08	3.70	3.95	3.80	3.65	3.75	3.50	3.57	4.00
B. Participle modif	2.50	3.40	3.76	1.20	1.70	2.33	2.17	0.86	2.33
C. Wh-- fronting	3.50	3.80	3.86	3.30	3.44	3.58	3.83	3.57	4.00
D. Reltvz / Claus conj	2.92	3.40	3.57	2.90	3.04	3.25	2.83	3.43	3.50
E. Reltvz / Pron delet	3.17	3.40	3.71	2.70	2.91	3.17	3.00	3.14	3.50
F. Double transform	3.75	3.50	3.95	3.90	3.83	3.75	3.83	3.71	3.83
G. Ask (q) / Tell	2.92	3.50	3.57	3.10	2.70	3.17	2.83	3.14	3.17
H. Easy to see	3.83	4.00	3.81	3.30	2.61	3.08	3.00	2.00	4.00
J. Promise / Tell	2.75	3.30	3.33	2.60	3.09	3.50	2.67	2.86	2.83
K. Indirect speech	3.42	3.70	3.71	3.40	3.65	3.42	3.17	3.57	3.50
L. Ask (r) / Tell	3.75	3.80	3.71	3.80	3.04	3.58	3.50	3.29	4.00
M. Pseudoimperatives	3.58	3.70	3.86	3.50	3.09	3.58	3.50	3.71	3.67
P. Pron ref	1.81	3.26	3.86	3.26	3.11	3.10	2.10	2.69	3.05
R. Pron ref (MDP)	1.27	3.28	3.73	3.04	3.17	2.60	1.60	2.97	2.13
S. Nominal subst	1.47	3.36	3.73	1.84	2.85	2.60	2.13	2.74	1.87
T. Clausal subst	1.22	2.27	3.05	2.67	3.01	3.22	0.89	1.90	2.67

(Scores on Subtasks: P, R, S, T are converted to an equivalent of a maximum of 4.00 to coincide with the scale for all other subtasks.)

Table 11: Task 1 and Task 2 Subtask mean scores, Language group by Grade interval

Hypothesis II : Discussion

Appreciation of contrasts across grade intervals may be gained by examining the patterns of subtask mean scores for Task 1 and Task 2 (Table 11). Departures from a curve of augmented growth in any of these skills of linguistic interpretation are in most cases minimal, particularly for native speakers. The more frequent interruptions in the curves for the E.S.L. samples may be due to the limitations of sample size. Barring the possible effects of differences in familiarity, there is no conceptual reason why older learners should find particular structures more difficult to comprehend than do younger pupils. Discussion with participating teachers offered no information to suggest the content of instruction experienced by any one grade interval differed from that of the other intervals within the E.S.L. samples respecting the variables examined. Insofar as possible, Ss were drawn from a number of programmes so as to randomize such instructional effects. These factors, combined with the observation that deviations from a positive curve occur with nearly equal frequency at the lower and upper ends of the grade interval range, tend to suggest the probability of fewer instances of such violations if the sample size of E.S.L. students were increased.

The anaphora subtasks evidence a pattern that substantiates the Bormuth et al. (1970) claim that school age children are unable to comprehend many common syntactic structures. This is reflected in the large gains between elementary and junior secondary native speakers, an observation that can also be generally applied to the E.S.L. samples where the gradient is less steep, possibly as a result of differences in curriculum content for native and non-native pupils in the elementary grades. These substantial gains by secondary students also support the contention that

Grade 4 pupils' capacity for anaphora resolution may be constrained by short term memory limitations (Lesgold, 1972).

In discussing results related to the preceding hypothesis, it was suggested that anaphora structures are often featured in E.S.L. curricula. If E.S.L. teachers do consciously identify this instructional need, it appears that there is a tendency to bring non-native students of all ages to a fairly uniform criterion, evidenced by the narrower range of anaphora subtask mean scores for E.S.L. students across grade intervals in contrast to the broad developmental gradient exhibited by native speakers. Particularly, senior secondary E.S.L. students seem to be in need of more instruction in anaphoric interpretation. Not only do they lag behind junior secondary students, but this is the only interval at which the E.S.L. samples perform significantly below native speaker counterparts.

Regarding the measures of text comprehension -- Tasks 3, 4, and 5, it is perhaps interesting to note that, for native speakers, significant differences are found in the narrow range between junior secondary and senior secondary intervals but not between the elementary and junior secondary. In contrast, where a significant difference occurs on any of these three tasks between E.S.L. intervals, it is only for a broad range comparison between elementary and senior secondary intervals. These observations, of course, need to be repeated on larger samples before the point of major gain in these skills can be designated with a reasonable degree of certainty.

Hypothesis III : Results

The rank order of difficulty of the sixteen subtasks of Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents) as determined by subtask mean scores does not vary among any subsamples.

The appropriate statistic to test this hypothesis is Kendall's coefficient of concordance (W). Total agreement among a group of three or more subsamples is manifest if W = 1.000; a maximum disarray is expressed by W = .000.

An inspection of Table 11 shows a few instances wherein subtask mean scores differ by less than .05. It was considered that this narrow margin was insufficient to merit separate rankings; rather, all subtask mean scores were rounded to one decimal place to treat small variations as tied rankings before computing the coefficients which, in each case, contain a correction for ties. An overall comparison for nine subsamples resulted in W = .663 ($p < .001$).

Additional comparisons were made by language group and by grade interval and are here summarized.

N.S.	E.S.L.(A)	E.S.L.(B)
.736	.780	.845

Table 12: Kendall's coefficient of concordance by Language group ($p < .01$)

N.S.	E.S.L.(A)	E.S.L.(B)
.859	.620	.681

Table 13: Kendall's coefficient of concordance by Grade interval ($p < .025$)

These results clearly indicate that even the most consistent grouping falls short of total agreement.

Hypothesis III : Conclusions

The hypothesis of no difference in the rank order of subtask difficulty among subsamples cannot be accepted as significant coefficients demonstrate similarities among subsamples to fluctuate within a moderate to high range of correlation.

Hypothesis III : Discussion

Two findings of some interest arise from the test of this hypothesis based on the assumption that relative difficulty of linguistic tasks should remain constant for both first and second language learners and at varying stages of language development.

The first is an indication of a somewhat greater degree of similarity across grade intervals among E.S.L. students than for native speakers (Table 12). One possible reason for this phenomenon may be that the conscious English language experience of the non-native student is of necessity more controlled through a narrow focus of E.S.L. instruction in contrast to the greater variety of language forms to which the native speaker can attend.

The second observation to be made is that the coefficients across language groups are lower for secondary than for elementary students (Table 13). In retrospect, from a developmental view, this trend is to be expected as native speakers have, over several years of growth, increasing opportunity to diversify their language experience while more mature E.S.L. students might be selective in acquiring those syntactic structures perceived to most adequately meet their linguistic

needs. Whatever the cause of this pattern, an important instructional consideration is implicit in the results. Based on the sampling criteria employed for this study, it would appear that elementary E.S.L. students about to be admitted to integrated (native and non-native speaker) classes in which a full educational programme is offered are more similar to native speaker peers in the nature of their language competence as measured by syntactic comprehension than their counterparts at the secondary grade intervals.

Comparison of samples by Language group

In accord with the plan of this study as outlined in Chapter IV, the grade intervals of each sample -- native speakers N.S. , Chinese learners of English E.S.L.(A) , and E.S.L. students of other native languages E.S.L.(B) -- are next combined to form three single groups.

Consideration of the number of Ss at each grade interval in the combined samples (Table 14), however, suggested that the three groups may not be comparable. While a symmetry is apparent in the E.S.L. samples wherein one half of the Ss are at the junior secondary interval or below and the other half are at that interval or above, it was noted that the native speaker sample is skewed by a greater number of Ss at the senior secondary interval.

	Elem	JrSec	SrSec
N.S.	12	10	21
E.S.L.(A)	10	23	12
E.S.L.(B)	6	7	6

Table 14: Number of Ss in each Language group by Grade interval

To adjust for this bias, nine Ss at this interval were eliminated from the collective native speaker sample through a table of random numbers in order to gain a symmetry across grade intervals similar to that of the E.S.L. samples (Table 14a).

Elem	JrSec	SrSec
12	10	12

Table 14a: Number of Ss in adjusted Native Speaker sample by Grade interval

A check on the data pertaining to these combined subsamples indicates that the requisites for the parametric analysis of variance procedure -- a reasonable approximation to normality and homogeneity of variances -- have been generally satisfied.

Hypothesis IV : Results

Combined across grade intervals,

(1) the performance of native speakers is superior to that of students for whom English is a second language, and

(2) the performance of Chinese speakers does not differ from that of other E.S.L. students

on each task and subtask included in the present instrument.

A one-way analysis of variance was conducted on each set of task and subtask scores. 'A posteriori' multiple range tests were selected to locate significant differences among sample mean scores: (1) Scheffe (alpha = .05), and (2) Duncan (alpha = .05). These two tests may be viewed as complementary checks against Type I and Type II errors, respectively,

i.e., at a specific alpha level, the Scheffe procedure yields the fewest significant differences, the Duncan, the most, when applied to the same data base (Ferguson, 1976:300). In each instance where a significant difference is claimed, it is substantiated by both procedures with one notable exception for Task 2.

Results depicted in Tables 15 - 19 may be summarized.

Task 1 as a whole significantly discriminates between native speakers and both groups of E.S.L. students. Associated subtasks contributing to this result are B: Participle modifiers and H: Easy to see, and, to a lesser extent as they only separate the Chinese speakers from the other two groups, E: Relativization by pronoun deletion and M: Pseudo-imperatives.

Task 2 results are less certain, there being a discrepancy between the two multiple range procedures: the Duncan (.05) test finds a significant difference between Chinese and the other E.S.L. students; the Scheffe (.05) appraises all three groups to be one homogeneous subset. To resolve this conflict, an additional procedure, Student-Newman-Keuls (.05), described as a "compromise" measure between Type I and Type II errors (Ferguson, ibid.), was conducted, leading to no significant differences. This indication and the observation that the one-way analysis of variance for Task 2 yields an insignificant F ratio ($p = .111$) prompts the investigator to conclude that Task 2 as a whole cannot be said to discriminate among any of the language groups. One associated subtask, T: Clausal substitution, nevertheless reveals a clearly superior performance by Chinese speakers over that of either the other E.S.L. students or native speakers.

Comparison of samples by Language group

The following sequence is used in presenting summary comparisons of the performance of subsamples combined across Grade intervals:

(1) Mean scores, variances, and number of Ss, listed in this format:

42.32 ← Subsample mean score

34 24.41 ← Subsample variance in scores

↑ Number of Ss in the subsample

(2) Indices of homogeneity of variances

(3) Multiple range tests

---

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
42.32		37.82		38.89	
34	24.41	45	19.83	19	28.43

$F = 8.80\frac{1}{2}$, $p < .01$

(2) Maximum s^2 / Minimum s^2 = 1.434

Bartlett - Box F = 0.458, not sig.

(3)

Scheffe (.05)	Duncan (.05)
N.S. v. E.S.L.(A),E.S.L.(B)	N.S. v. E.S.L.(A),E.S.L.(B)

Table 15: Comparison of performance on Task 1 by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.56		3.71		3.68	
34	0.678	45	0.301	19	0.228

F = 0.564, not sig.

(2) Maximum s^2 / Minimum s^2 = 2.974

Bartlett - Box F = 4.642, $p < .01$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15a: Comparison of performance on Subtask A by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.18		1.76		1.74	
34	1.059	45	1.462	19	1.982

F = 15.906, $p < .001$

(2) Maximum s^2 / Minimum s^2 = 1.872

Bartlett - Box F = 1.197, not sig.

(3)

Scheffe (.05)	Duncan (.05)
N.S. v. E.S.L.(A),E.S.L.(B)	N.S. v. E.S.L.(A),E.S.L.(B)

Table 15b: Comparison of performance on Subtask B by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.68		3.44		3.79	
34	0.407	45	0.571	19	0.287

F = 2.134, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.992

Bartlett - Box F = 1.509, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15c: Comparison of performance on Subtask C by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.29		3.07		3.26	
34	0.578	45	0.245	19	0.649

F = 1.316, not sig.

(2) Maximum s^2 / Minimum s^2 = 2.645

Bartlett - Box F = 4.489, $p < .025$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15d: Comparison of performance on Subtask D by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.41		2.93		3.21	
34	0.613	45	0.427	19	0.509

F = 4.446, $p < .025$

(2) Maximum s^2 / Minimum s^2 = 1.435

Bartlett - Box F = 0.613, not sig.

(3)

Scheffe (.05)	Duncan (.05)
N.S. v. E.S.L.(A)	N.S. v. E.S.L.(A)

Table 15e: Comparison of performance on Subtask E by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.76		3.82		3.79	
34	0.185	45	0.149	19	0.175

F = 0.195, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.240

Bartlett - Box F = 0.231, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15f: Comparison of performance on Subtask F by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.35		2.91		3.05	
34	1.205	45	0.992	19	0.941

F = 1.807, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.280

Bartlett - Box F = 0.245, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15g: Comparison of performance on Subtask G by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.82		2.89		2.95	
34	0.210	45	0.874	19	0.941

F = 14.272, $p < .001$

(2) Maximum s^2 / Minimum s^2 = 4.476

Bartlett - Box F = 9.082, $p < .001$

(3)

Scheffe (.05)	Duncan (.05)
N.S. v. E.S.L.(A),E.S.L.(B)	N.S. v. E.S.L.(A),E.S.L.(B)

Table 15h: Comparison of performance on Subtask H by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.24		3.09		2.79	
34	0.731	45	0.855	19	1.398

F = 1.326, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.912

Bartlett - Box F = 1.349, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15i: Comparison of performance on Subtask J by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.59		3.53		3.42	
34	0.431	45	0.436	19	0.591

F = 0.368, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.369

Bartlett - Box F = 0.362, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15j: Comparison of performance on Subtask K by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.71		3.36		3.58	
34	0.456	45	1.098	19	0.591

F = 1.578, not sig.

(2) Maximum s^2 / Minimum s^2 = 2.406

Bartlett - Box F = 3.703, $p < .025$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 15k: Comparison of performance on Subtask L by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.74		3.31		3.63	
34	0.261	45	0.537	19	0.690

F = 4.014, $p < .025$

(2) Maximum s^2 / Minimum s^2 = 2.642

Bartlett - Box F = 3.276, $p < .05$

(3)

Scheffe (.05)	Duncan (.05)
N.S. v. E.S.L.(A)	N.S. v. E.S.L.(A)

Table 151: Comparison of performance on Subtask M by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
13.56		14.67		11.63	
34	48.14	45	12.41	19	26.69

F = 2.253, not sig.

(2) Maximum s^2 / Minimum s^2 = 3.879

Bartlett - Box F = 8.452, $p < .001$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	E.S.L.(A) v. E.S.L.(B)

Table 16: Comparison of performance on Task 2 by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
5.24		5.49		4.58	
34	6.246	45	1.710	19	3.702

F = 1.511, not sig.

(2) Maximum s^2 / Minimum s^2 = 3.652

Bartlett - Box F = 7.740, $p < .001$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 16a: Comparison of performance on Subtask P by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.41		3.73		2.84	
34	3.765	45	1.109	19	3.362

F = 2.173, not sig.

(2) Maximum s^2 / Minimum s^2 = 3.394

Bartlett - Box F = 7.547, $p < .001$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 16b: Comparison of performance on Subtask R by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
3.47		3.20		2.84	
34	3.590	45	2.073	19	2.474

F = 0.909, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.732

Bartlett - Box F = 1.451, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 16c: Comparison of performance on Subtask S by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
1.59		2.24		1.37	
34	1.462	45	0.598	19	1.023

F = 7.019, p .01

(2) Maximum s^2 / Minimum s^2 = 2.444

Bartlett - Box F = 3.754, $p < .025$

(3)

Scheffe (.05)	Duncan (.05)
E.S.L.(A) v. N.S.,E.S.L.(B)	E.S.L.(A) v. N.S.,E.S.L.(B)

Table 16d: Comparison of performance on Subtask T by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
15.06		16.39		15.78	
31	20.39	44	7.31	18	12.18

F = 1.264, not sig.

(2) Maximum s^2 / Minimum s^2 = 2.789

Bartlett - Box F = 4.659, $p < .01$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 17: Comparison of performance on Task 3 by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
16.68		16.29		15.47	
34	13.80	45	8.03	19	12.71

F = 0.809, not sig.

(2) Maximum s^2 / Minimum s^2 = 1.719

Bartlett - Box F = 1.516, not sig.

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 18: Comparison of performance on Task 4 by Language group

(1)

N.S.		E.S.L.(A)		E.S.L.(B)	
14.13		14.59		13.94	
31	20.78	44	8.06	18	15.82

F = 0.250, not sig.

(2) Maximum s^2 / Minimum s^2 = 2.578

Bartlett - Box F = 4.080, $p < .025$

(3)

Scheffe (.05)	Duncan (.05)
no sig. diff.	no sig. diff.

Table 19: Comparison of performance on Task 5 by Language group

The measures of text comprehension -- Tasks 3, 4, and 5, are all characterized by no significant differences across language groups.

Considering each language group in turn, native speakers significantly outperform both groups of E.S.L. students on Task 1 and associated subtasks, B: Participle modifiers and H: Easy to see, only. On an additional two subtasks, E: Relativization by pronoun deletion and M: Pseudoimperatives, native speaker performance is significantly superior to that of Chinese speakers.

Chinese speakers significantly outperform native speakers as well as other E.S.L. students on Subtask T: Clausal substitution of Task 2.

Non-Chinese E.S.L. students in no instance significantly outperform either of the other two groups.

Hypothesis IV : Conclusions

The hypothesis of native speaker superiority cannot be established by the present results, except for Task 1, and by the failure to attain significant differences on twelve out of sixteen subtasks.

The hypothesis of no difference between the E.S.L. groups is supported by the results of all tasks, despite any question concerning Task 2, and by the patterning of scores on fifteen of the sixteen subtasks.

Hypothesis IV : Discussion

As stated earlier, even after a lengthy period of instruction, E.S.L. students remain generally inferior to native speakers in their comprehension of most syntactic forms tested. Only in a few instances, notably in connection with passivization transforms, do E.S.L. mean scores exceed those of the native speaker sample on subtasks of grammatical implication.

A review of the relativization subtasks: C, D, and E, shows Chinese speakers to be at a noticeable disadvantage from other E.S.L. students who compare rather favourably with the native speaker sample. For both groups of E.S.L. learners, however, certain of the items indicate a lack of facility in interpreting "nested" forms -- e.g., The boy (the girl hit) fell down.

Similarly, Chinese speakers seem to have more difficulty than other learners in discerning contrasts in sentence meaning that involve the minimal distance principle, although the other E.S.L. group also consistently scores below the native speaker sample on the four related subtasks: G, H, J, and L.

Among the twelve subtasks of Task 1, two stand out as effective discriminators between native speakers and either of the E.S.L. groups -- B: Participle modifiers and H: Easy to see. Familiarity with E.S.L. curriculum materials will readily suggest why this should be so. Neither structure is given much emphasis in basic programmes, especially in terms of deliberately manipulating the syntactic patterns to illustrate contrasts in meaning. Subtask M: Pseudoimperatives discriminates only against the Chinese learners, and to a lesser extent than the aforementioned two, even though it also tests a seldom taught construction.

Subtask N: Agentless passivization is of some interest in that the great majority of Ss, native speakers and E.S.L. students alike, may have given too broad an interpretation to the lexeme, "someone", allowing the inclusion of non-human referents. Preliminary item analysis data indicated a disparity between this and the other subtasks of Task 1 for which reason it was removed from all subsequent analyses in the study. Consequently, no comparative mean scores are reported.

Comparative data for the anaphora subtasks are counter to initial expectations. While native speaker mean scores are higher than those for the multiethnic E.S.L. group on all subtasks, they fall below those attained by the Chinese speakers on three of the four subtasks.

It is difficult to postulate a reason for these results. Ss for both E.S.L. samples were drawn from the same classes thereby suggesting equality of instructional opportunity. While the anaphora subtasks were cast in a "constructed response" format, the investigator endeavoured to apply consistent standards of acceptability to the responses of all Ss. A contrast in language typology does not readily account for the superior performance of the Chinese learners nor can it explain the inferior scores of the native speakers.

The lack of significant differences among the three samples on the measures of text comprehension -- Tasks 3, 4, and 5 has already been observed in comments on Hypothesis I where some of the implications of this finding were noted.

Hypothesis V : Results

> The rank order of difficulty of the sixteen subtasks of Task 1 (recognition of grammatical implication) and Task 2(identification of anaphoric referents) as determined by subtask mean scores does not vary among any subsamples combined across grade intervals.

After adjusting subtask mean scores listed in Table 20 by rounding to one decimal place to reduce the effect of numerous small, and possibly spurious, differences in rankings, Kendall's coefficient of concordance, corrected for ties, was computed upon the three language groups, resulting in $W = .809$ ($p < .01$).

N. S.		E.S.L.(A)		E.S.L.(B)	
H. Easy to see	3.82	F. Double transform	3.82	C. Wh-- fronting	3.79
F. Double transform	3.76	A. Passivization	3.71	F. Double transform	3.79
M. Pseudoimperatives	3.74	K. Indirect speech	3.53	A. Passivization	3.68
L. Ask (r) / Tell	3.71	C. Wh-- fronting	3.44	M. Pseudoimperatives	3.63
C. Wh-- fronting	3.68	L. Ask (r) / Tell	3.36	L. Ask (r) / Tell	3.58
K. Indirect speech	3.59	M. Pseudoimperatives	3.31	K. Indirect speech	3.42
A. Passivization	3.56	P. Pron ref	3.14	D. Reltvz / Claus conj	3.26
E. Reltvz / Pron delet	3.41	J. Promise / Tell	3.09	E. Reltvz / Pron delet	3.21
G. Ask (q) / Tell	3.35	D. Reltvz / Claus conj	3.07	G. Ask (q) / Tell	3.05
D. Reltvz / Claus conj	3.29	T. Clausal subst	2.99	H. Easy to see	2.95
J. Promise / Tell	3.24	R. Pron ref (MDP)	2.98	J. Promise / Tell	2.79
B. Participle modif	3.18	E. Reltvz / Pron delet	2.93	P. Pron ref	2.62
P. Pron ref	2.99	G. Ask (q) / Tell	2.91	R. Pron ref (MDP)	2.27
S. Nominal subst	2.78	H. Easy to see	2.89	S. Nominal subst	2.27
R. Pron ref (MDP)	2.73	S. Nominal subst	2.56	T. Clausal subst	1.83
T. Clausal subst	2.12	B. Participle modif	1.76	B. Participle modif	1.74

(Scores on Subtasks: P, R, S, T are converted to an equivalent of a maximum of 4.00 to coincide with the scale for all other subtasks.)

Table 20: Comparative rankings of subtask difficulty by Language group

Of further interest is the degree of similarity in rank order of subtask difficulty between the two E.S.L. samples and between each E.S.L. sample and the native speaker sample. Kendall's tau, a non-parametric correlation coefficient, was selected as the index most suited to comparisons containing several tied ranks (cf. Nie et al., 1975:289). Results, corrected for ties, are summarized.

E.S.L.(A)	with	E.S.L.(B)	.655
E.S.L.(A)	with	N.S.	.382
E.S.L.(B)	with	N.S.	.687

Table 21: Kendall tau coefficients for paired comparisons among Language groups ($p < .025$)

Hypothesis V : Conclusions

While it appears that the three language groups possess a rather high degree of agreement in their rankings of subtask difficulty as indicated by the coefficient of concordance, a closer examination of the data utilizing a more conservative measure, Kendall's tau, yields more moderate correlations between the two E.S.L. groups and the multiethnic students [E.S.L.(B)] and the native speakers. By contrasts, Chinese learners [E.S.L.(A)] show a much lower degree of agreement with native speakers on the relative difficulty of subtasks. Therefore, the hypothesis of no difference among the combined subsamples cannot be accepted.

Hypothesis V : Discussion

Having established a degree of disparity in the rankings between Chinese and non-Chinese learners (tau = .655), it is of some interest to note that the two E.S.L. groups, generally comparable in their overall

performance on the subtasks[4], show such marked differences in their degrees of agreement with the rank order of subtask difficulty established for native speakers.

The procedure of paired comparisons was undertaken to provide some indication as to whether or not a specific native language determines which syntactic forms of English are the most difficult to comprehend. The striking contrast in coefficients obtained by comparisons with native speakers (tau = .382 and .687 for Chinese and non-Chinese students, respectively) does strongly suggest that the order of difficulty of transformations and anaphoric types tested is not independent of the students' native language. As might be expected, the group of which half are speakers of European languages performs in a way more similar to native speakers of English than do Chinese students of equal proficiency.

The remainder of this discussion will examine some notable trends among the specific subtasks.

Because the present study was not conceived upon any norms of linguistic familiarity for specified populations, no definitive statement can be made regarding the effect of this factor as discussed earlier in the literature review. With respect to transformational complexity, however, attention can be drawn to a number of interesting observations.

O'Donnell, Griffin, and Norris (1967) report that younger children use relative clauses much more frequently than noun modification by a participle. The present results indicate that all three language groups find

---

4. Of the 68 items comprising the sixteen subtasks, the E.S.L.(A) sample mean is 52.49 (77.2%); the E.S.L.(B) sample mean is 50.53 (74.3%). t = 0.94, not sig.

structures containing a participle as a noun modifier (Subtask B) more difficult to interpret than any of the forms of relativization tested.

Another expectation that is confirmed supposes a subtask that test understanding of two structures, both of which conform to the minimal distance principle, is easier than a subtask involving a contrast between a conforming and a non-conforming structure. As indicated in Table 20, Subtask L: Ask (request) / Tell elicits a better performance from all groups of Ss than either Subtask G: Ask (query) / Tell or Subtask J: Promise / Tell.

Lastly, from the viewpoint of transformational complexity, it might be predicted that Subtask F involving both relativization and passivization would be more difficult than subtasks characterized by only one of these transformations, i.e., Subtasks A, D, and E. Contrary to such expectations, the predicted order of difficulty is violated by all groups. This result may be explained partly by the presence in most items of Subtasks D and E of relative pronoun deletion, a transform which has been reported to be especially difficult for elementary pupils (Fagan, 1969).

One other prediction, also unconfirmed, is that Subtask M: Pseudoimperatives would be among the most difficult for the E.S.L. samples. Even if the somewhat idiomatic usage of these structures precludes them from an E.S.L. curriculum, this subtask ranks as fairly easy for both native and non-native speakers.

Performance on the anaphora subtasks must be interpreted with caution; it can be argued that scores for Subtasks P, R, S, and T are not directly comparable to those of other subtasks since a "constructed response" format was employed for the former, a "yes/no choice" for the latter; hence,

the tendency for anaphora subtask scores to filter toward the bottom ranks. If a "yes/no choice" format does generally facilitate performance over "constructed responses", the marked difficulty of Subtask B: Participle modifiers for E.S.L. students becomes all the more apparent.

Finally, these four subtasks were thought to represent a hierarchy of complexity, particularly in the context of the S having to form his own response. The predicted order of subtask mean scores was: P: Pronominal reference R: Pronominal reference (Minimal Distance Principle) S: Nominal substitution T: Clausal substitution. Results, however, do not reflect this pattern across the three language groups. The contrast in difficulty between Subtask P and Subtask R is nevertheless supported.

Hypothesis VI : Results

> A comparison of the three measures of text comprehension exhibits a pattern of diminished performance wherein Task 4 (judgment of truth value) scores are greater than Task 3 (indication of textual locus) scores which are greater than Task 5 scores.

This hypothesis was tested by means of a series of correlated t-tests, results of which are presented in Tables 22a, b, c.

Hypothesis VI : Conclusions

While the hypothesis is supported by the data for native speakers, differences between Task 3 and Task 4 performance do not occur as predicted for either of the E.S.L. samples.

It is also to be noted that, for each group, highly significant differences are evident between each of the component tasks and the combined measure (Task 5), thereby indicating that neither Task 3 nor Task 4 can

	Mean	st dev	t value	
Task 4	17.00	3.45		
Task 3	15.06	4.52	4.76	$p < .001$
Task 5	14.13	4.56	5.40	$p < .001$

* Task 4 / Task 5, t value = 8.22, $p < .001$

Table 22a: Correlated t-tests for Native Speakers (N = 31) across three measures of text comprehension

	Mean	st dev	t value	
Task 4	16.39	2.79		
Task 3	16.39	2.70	0.00	not sig.
Task 5	14.59	2.84	7.69	$p < .001$

* Task 4 / Task 5, t value = 10.92, $p < .001$

Table 22b: Correlated t-tests for E.S.L.(A) Students (N = 44) across three measures of text comprehension

	Mean	st dev	t value	
Task 4	15.50	3.67		
Task 3	15.78	3.49	-0.57	not sig.
Task 5	13.94	3.98	7.46	$p < .001$

* Task 4 / Task 5, t value = 4.18, $p < .001$

Table 22c: Correlated t-tests for E.S.L.(B) Students (N = 18) across three measures of text comprehension

stand alone as an adequate measure of comprehension as earlier defined for this study.

Hypothesis VI : Discussion

The suggested rationale for the predicted order of difficulty for the three measures of text comprehension was based on relative probabilities of correct responses. Where the predicted order is upheld, however, alternative explanations are possible; where results contradict probabilities, other accounts must be sought.

Native speaker performance observes the predicted pattern. One reason Task 3 (indication of textual locus) scores are lower than those for Task 4 (judgment of truth value) may be that the Ss, particularly at the secondary intervals, found the passage relatively easy to comprehend on first reading. If so, information storage, be it accurate or not, would have been facilitated, rendering the task of citing the locus for one's response unusual and bothersome.

By contrast, the E.S.L. samples appear to have found one component task as difficult as the other. It may be that second language learner's comprehension strategies are less global than those of native speakers. Consequently, they would be somewhat accustomed to scruitinizing a text for individual sentences that provide information appropriate to a given item statement such as appears in Test No. 4.

The foregoing explanations are admittedly speculative. Because the administration of the instrument did not call for the enforcement of a strict time limit, it is not possible to present evidence here to suggest that a higher incidence of response verification through locus searching did occur among the E.S.L. samples.

As an additional note of interest, Frase and Washington (1970)

found that, for elementary school pupils, comprehension errors increase sharply if more than one sentence has to be processed. A comparison of the item difficulties of eleven single sentence locus items (Nos. 1, 3, 5, 7, 11, 12, 16, 17, 18, 20, 21) with six multiple sentence locus items (Nos. 2, 4, 9, 13, 14, 22) -- q.v. Table 23 -- establishes that while for the entire sample of this study indicating the locus of multiple sentence items is noticably more exacting a task, judging their truth value is not much more difficult than for single sentence items.

	single sentence	multiple sentence
Task 3	.781	.445
Task 4	.755	.684
Task 5	.679	.426

Table 23: Mean item difficulties of single sentence and multiple sentence items for entire sample population (N = 107)

Hypothesis VII : Results

Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents) are each better predictors of Task 5 (the combined measure of locus indication and truth value judgment) than of Task 4 (judgment of truth value only). Task 1 and Task 2 are equally good predictors of Task 5 scores.

Correlations indicate that Task 1 and Task 2 are each better predictors of Task 5 than of Task 4 only for the native speaker sample. In the case of the E.S.L.(A) sample, Task 2, but not Task 1, better predicts Task 5 than Task 4 for these Chinese speakers. Neither Task 1 nor Task 2 predicts Task 5 better than Task 4 for the multiethnic E.S.L.(B) sample. (cf. Tables 24a, b, c)

	Task 1	Task 2	Task 3	Task 4	Task 5
Task 1	----	.581	.713	.661	.703
Task 2		----	.694	.664	.709
Task 3			----	.872	.977
Task 4				----	.919
Task 5					----

Table 24a: Correlation coefficients for Native Speakers completing all tasks (N = 31, $p < .01$)

	Task 1	Task 2	Task 3	Task 4	Task 5
Task 1	----	.481	.275	.417	.390
Task 2		----	.380	.598	.625
Task 3			----	.797	.845
Task 4				----	.925
Task 5					----

Table 24b: Correlation coefficients for E.S.L.(A) Students completing all tasks (N = 44, $p < .01$)

	Task 1	Task 2	Task 3	Task 4	Task 5
Task 1	----	.484	.436	.540	.451
Task 2		----	.255	.201	.189
Task 3			----	.837	.969
Task 4				----	.918
Task 5					----

Table 24c: Correlation coefficients for E.S.L.(B) Students completing all tasks (N = 18, $p < .01$)

Correlation coefficients between each predictor and the criterion are nearly equal in the case of native speakers. Data pertaining to second language students identifies Task 2 as a better predictor than Task 1 for the E.S.L.(A) sample, the converse being true for the E.S.L.(B) sample.

Zero order correlations, however, do not explain the full significance of the relationship of each task to the criterion measure of comprehension. In order to adequately assess the relative importance of the linguistic skills tested by each task to the act of text comprehension as defined at the outset of this study, a more complex statistical procedure known as multiple regression analysis is required.

Multiple regression analyses of two sorts are applied to the present data. Type I sets an identical inclusion level (Nie et al., 1975:344) for all independent variables -- in this case, Task 1 and Task 2 -- to be considered in any one analysis. This means that all independent variables enter the analysis on an equal footing. The practical effect of this procedure is to first identify the leading independent variable and report the amount of variance it contributes to the dependent variable, Task 5. This variance is referred to herein as "independent variance" since it comprises that variance uniquely contributed by the leading variable plus any proportion of variance contributed in common with the other independent variable. In other words, the "independent variance" associated with a given task is that proportion of variance in Task 5 scores that can be explained without reference to any other task. It is the same as would result if only that particular task were entered into the regression equation. Next, the less crucial independent variable is identified and the amount of additional variance it contributes to the dependent criterion measure is reported.

Type I			Type II		
Task 2	+	(Task 1)	Task 1	+	(Task 2)
50.3%	+	(12.7%)	49.4%	+	(13.6%)

Total variance explained = 63.0%

Table 25a: Multiple Regression Analyses: independent variance and (additional variance) contributed by syntactic measures to criterion Task 5 scores of Native Speakers (N = 31, $p < .01$)

Type I			Type II		
Task 2	+	(Task 1)	Task 1	+	(Task 2)
39.1%	+	(1.0%)	15.2%	+	(24.9%)

Total variance explained = 40.1%

Table 25b: Multiple Regression Analyses: independent variance and (additional variance) contributed by syntactic measures to criterion Task 5 scores of E.S.L.(A) Students (N = 44, $p < .01$)

Type I			Type II		
Task 1	+	(Task 2)	Task 2	+	(Task 1)
20.4%	+	(0.1%)	3.6%	+	(16.9%)

Total variance explained = 20.5%

Table 25c: Multiple Regression Analyses: independent variance and (additional variance) contributed by syntactic measures to criterion Task 5 scores of E.S.L.(B) Students (N = 18, not sig.)

In contrast to the foregoing procedure, Type II analyses specify the sequence in which the independent variables enter the analysis thereby establishing a hierarchy of inclusion levels. This procedure was conducted subsequent to obtaining results for Type I analyses in order to assess the total amount of variance contributed by the less crucial task.

The results of separate analyses for each language group are summarized in Tables 25a, b, c.

Hypothesis VII : Conclusions

The hypothesis that the linguistic skills embodied in the first two tasks better predict Task 5 than Task 4 scores is supported only in the case of native speakers, the data from the E.S.L. samples bringing mixed or contrary results.

While Task 1 and Task 2 appear to be equally good predictors of Task 5 scores for native speakers, this does not prove to be true among E.S.L. students where one task predicts better than the other depending on the particular sample.

Marked differences in the amount of variance explained by Task 1 and Task 2 are noted across language groups. In particular, native speaker performance on Task 5 can be more adequately accounted for by reference to these two tasks than can be the variance in E.S.L. students' scores on the same criterion measure. It is to be noted, however, that the regression coefficients for the E.S.L.(B) sample are not statistically significant.

Hypothesis VII : Discussion

In order to better interpret the foregoing results, it should be recalled that it has already been demonstrated that there is no significant difference among the three language groups in their performance on the

criterion Task 5. This fact makes it possible to conduct comparisons among samples of equal proficiency to determine the extent to which various populations are able to utilize the skills measured by the first two tasks to perform the criterion task. Only through such a componential examination of language behaviour can the primary purpose of this study be accomplished.

Results indicate that where significant differences occur among groups on the syntactic measures, a greater proportion of variance in the criterion scores of the superior group can be explained by those variables. Specifically, Task 1 results discriminate between native speakers and the E.S.L. samples. It is observed that the variance contributed by Task 1 to Task 5 scores is more than twice as great for native speakers than for either of the E.S.L. samples. While the investigator prefers to consider that no significant differences are present among the three language groups on Task 2, one multiple range test indicates that the two E.S.L. samples are not from the same subset. In this case, the additional variance contributed by Task 2 to Task 5 scores of the superior E.S.L.(A) group is probably far greater than that for the E.S.L.(B) sample even though the regression coefficients for the latter are not significant.

In summation, it appears that different populations utilize the linguistic skills tested to various extents in accord with their demonstrated ability in those skills. Of greater consequence, however, is the finding that while the capacity to recognize grammatical implication relations and identify anaphoric referents is advantageous, the two skills accounting for most of the variance in the criterion scores of the ablest group, such ability is not crucial to success on what have been proposed to be requisites for demonstrating sentence comprehension within a text,

i.e., indication of textual locus and judgment of truth value.

This last outcome may be better appreciated by bearing in mind that variations across populations in the importance of syntactic variables are commonly incurred in language research. For example, Takahashi (1975) found differences between Grade 6 and Grade 9 pupils in the extent to which syntactic comprehension affected scores on a standardized test of reading comprehension while Baines (1975) reports that syntactic complexity in written composition is not an equally good predictor of reading comprehension for elementary and secondary students.

Hypothesis VIII : Results

Task 3 (indication of textual locus) is a better predictor than Task 4 (judgment of truth value) of the criterion for comprehension, Task 5.

Zero order correlations for the three measures of text comprehension are included in Tables 24a, b, c. The results of the associated multiple regression analyses are summarized in Tables 26a, b, c.

Hypothesis VIII : Conclusions

The hypothesis that Task 3 better predicts Task 5 scores than does Task 4 is supported by the data from the native speaker and the multiethnic E.S.L. samples, having incurred contrary results for the Chinese speaking E.S.L. group.

Multiple regression analyses identify Task 3 as leading Task 4 in contributing to Task 5 scores for the native speaker and multiethnic E.S.L. samples while Task 4 explains more variance than Task 3 on the criterion measure for the Chinese speaking E.S.L. group.

Type I			Type II		
Task 3	+	(Task 4)	Task 4	+	(Task 3)
95.6%	+	(1.8%)	84.4%	+	(13.0%)

Total variance explained = 97.4%

Table 26a: Multiple Regression Analyses: independent variance and (additional variance) contributed by component measures to criterion Task 5 scores of Native Speakers (N = 31, $p < .01$)

Type I			Type II		
Task 4	+	(Task 3)	Task 3	+	(Task 4)
85.6%	+	(3.2%)	71.4%	+	(17.4%)

Total variance explained = 88.8%

Table 26b: Multiple Regression Analyses: independent variance and (additional variance) contributed by component measures to criterion Task 5 scores of E.S.L.(A) Students (N = 44, $p < .01$)

Type I			Type II		
Task 3	+	(Task 4)	Task 4	+	(Task 3)
94.0%	+	(3.8%)	84.2%	+	(13.6%)

Total variance explained = 97.8%

Table 26c: Multiple Regression Analyses: independent variance and (additional variance) contributed by component measures to criterion Task 5 scores of E.S.L.(B) Students (N = 18, $p < .01$)

Hypothesis VIII : Discussion

Considering the cited inadequacies of Task 4 as a measure of comprehension, the high correlations obtained with the more rigorous Task 5 were unexpected. Despite the reservations raised earlier, the empirical results of this investigation find Task 4 a reasonably accurate predictor of what has been argued to be a conceptually adequate measure of sentence comprehension within a text. This outcome may be attributed to the selection criteria which precluded students observed to have difficulties in reading comprehension. Differences between Task 3 and Task 4 correlations with the criterion measure are probably more likely to occur among samples of less able students more prone to random response strategies.

While the results of the multiple regression analyses for two of the three groups would seem to offer evidence that Task 3 is far more consequential to Task 5 scores than is Task 4, an anticipated outcome, additional analyses revealed that most of the variance attributable to either Task 3 or Task 4 is in all probability common variance shared between these two variables.

Since the correlations for both Task 3 and Task 4 with the criterion Task 5 fall within a narrow range, the question arises as to which of the two component tasks is a better measure of text comprehension. This is directed to a practical preference for a single measure that does not require the combining of component subscores.

For design of a test of literal comprehension to be administered to a sampling of students from the three language populations represented in this study, a certain choice is not indicated by the present data. Choosing to use a measure of indicating textual locus only would seem to offer the advantage of greater ease and flexibility in constructing items; one merely

selects a sentence or two from the passage and applies the transformations or introduces the anaphoric structure on which the students are to be tested. Relatively little concern need be given to the plausibility of the item statement. The disadvantages of this format lie in its novelty to the examinees and some possible variation in criteria for scoring.

By contrast the more familiar measure of judgment of truth value only, although easier to score in a very consistent manner, usually requires more time to construct since item statements containing plausible but false propositions must be devised. In formats calling for a "cannot say" option, items must be included for which it is difficult, if not impossible, to state what aspect of language competence is being tested.

Precisely because of this last consideration, it is here proposed that since empirical findings offer no clear indication, conceptual considerations should govern the decision as to which component task, locus or truth value, to adopt as a single measure of sentence comprehension within a text. In this light, the argument that knowledge of specific linguistic variables might be more readily attributable to responses calling for indication of textual locus than to those based solely on a judgment of truth value remains unrefuted.

Supplementary analysis of the data

Throughout the analysis of the data in the present study, two indicators of concurrent validity were continually monitored: (1) superior performance of native speakers in comparison to E.S.L. students, and (2) for native speakers, augmented performance as a function of higher grade interval. Beyond such perfunctory checks, it is necessary to examine the possibility that inappropriate response strategies may have determined performance

before concluding that scores reflect competence in the linguistic variables under investigation.

Task 1

Three strategies are considered: (1) random responses, (2) responses based on a similarity of sentence length, and (3) responses based on similarity of surface structure.

Random responses occur when the S can neither interpret one or both sentences of a pair nor establish any plausible device to suggest whether or not two sentences may be semantically equivalent. In this instance, a S may consistently respond either "yes" or "no" to all items or he may set a pattern of responses -- e.g., yes, yes, no, yes, yes, no... In preparing the test booklets for mechanical scoring, no such patterns were observed. Further, since the key for Tests No. 1 and No. 2 calls for 23 "yes" and 25 "no" responses, mean scores resulting from totally random response choices should approach approximately 24.00. All 107 Ss participating in the study received scores higher than 24. The lowest mean score among the nine subsamples is 36.74. Probability of attaining this high a score on the basis of random response choices is low ($p < .125$).

Similarity of sentence length might conceivably influence a S's decision as to whether or not two sentences bear the same underlying meaning. Defining "similar length" as the instance wherein the two sentences of a pair do not differ by more than five typewritten spaces in overall length, the 48 items of Tests No. 1 and No. 2 were divided into two categories.

In the first category were placed sentence pairs whose similarity of graphic length might possibly be a clue to similarity of meaning. Included were sentence pairs of similar graphic length that do have the same meaning and sentence pairs of different graphic length that do not have the

same meaning. These two types form a set of 25 items: Test No. 1 -- 4, 7, 8, 11, 12, 15, 16, 17, 18, 19, 23, 24, 27, 28, 29, 30, 31, 32, 33, 35; Test No. 2 -- 2, 7, 8, 9, 12. The remaining 23 items comprise the second category in which reliance upon similarity and difference in graphic length would lead to an incorrect response.

A correlated t-test revealed that performance on items in the first category is superior to that for the second category. The difference is significant ($p < .025$) for Native Speakers: Senior Secondary and all grade intervals of the E.S.L.(A) sample, thereby suggesting the possibility that some Ss may have relied upon graphic length clues to determine their responses rather than any syntactic-semantic interpretation of the sentences.

A comparison of the performance of the nine subsamples on the set of items thought to offer some clue to similarity or difference of meaning, however, yields a pattern of results very similar to the findings reported for Tests No. 1 and No. 2 as a whole (Table 27). While similarity of graphic length between sentences may be a factor in enhancing performance, it is concluded not to be a crucial variable contributing to differences in mean scores among the subsamples. It is most probable that responses to items in Tests No. 1 and No. 2 are generally based on some attribute other than visual discrimination of sentence length.

	Elem	JrSec	SrSec
N.S.	83.3% (81.6%)	91.5% (90.0%)	94.8% (93.4%)
E.S.L.(A)	84.4% (78.1%)	79.0% (76.5%)	86.7% (83.7%)
E.S.L.(B)	80.0% (78.8%)	80.0% (76.8%)	88.7% (88.2%)

Table 27: Proportion of correct responses for graphic length clue items and (all items) by Language group and Grade interval

One purpose of the present study is to assess Ss' proficiency in recovering deep structure through an accurate recognition of the function of syntactic elements in a sentence. It is, therefore, important in establishing the validity of Tests No. 1 and No. 2 as a measure of recognition of grammatical implication that it be demonstrated that performance on the tests cannot be attributed to Ss' strategy of relying solely upon surface structure patterns to attain a correct response.

It is conceivable that some Ss might examine a pair of sentences to determine if content lexemes -- nouns, verbs, adjectives, and adverbs -- occur in the same sequential order in both sentences. If so, the sentences are judged to be semantically equivalent; otherwise, the S concludes that the two sentences differ in meaning. For example,

The girl is easy to see.
The girl is easily seen.

are deemed to be paraphrases of each other.

Another strategy based on the surface structure of the sentences would consider the addition of elements in one sentence to preclude its being a paraphrase of another. Applying this principle,

Mary told Jack to come here today.
Mary told Jack that she should come here today.

are judged to be semantically not equivalent.

To determine the prevalence of these two strategies, it is possible to divide the 36 items of Test No. 1 into two groups based on whether or not similarities and differences in surface structure provide superficial clues to deep structure identity between sentences. The following 21 items are thought to provide such clues: 3, 4, 5, 7, 11, 12, 15, 16, 18, 20, 21, 22, 23, 24, 26, 27, 29, 30, 32, 33, 35.

A correlated t-test revealed that on Native Speakers: Senior Secondary actually performed significantly better on those items containing superficial clues. This is the one sample whose native language and grade interval factors would indicate the least dependency on gleaning surface structure clues. A small difference favouring the superficial clue items for the E.S.L.(A): Senior Secondary subsample is not significant. Moreover, the pattern of results is very similar for both the "superficial clues" only and the entire set of items comprising Test No. 1 (Table 28). It is to be noted that all but the aforementioned subsample of E.S.L. students performed better, in some cases significantly so, on that set of items believed not to contain surface structure clues to underlying meaning. It may, therefore, be concluded that the results of this phase of the investigation do reflect the comparative competencies of the nine subsamples in recovering the deep structure of item sentences.

	Elem	JrSec	SrSec
N.S.	80.6% (78.9%)	89.5% (88.9%)	94.5% (93.1%)
E.S.L.(A)	73.8% (74.4%)	73.3% (74.9%)	82.9% (82.2%)
E.S.L.(B)	73.8% (76.8%)	72.8% (73.0%)	84.1% (86.6%)

Table 28: Proportion of correct responses for "superficial clue" items and (all items) by Language group and Grade interval

Task 2

The "constructed rsponse" format of Test No. 3 generally precludes a patterning of responses based upon false strategies. Since the correct anaphoric referents occurred in various positions within item sentences,

the use of positional clues could not be expected to augment scores. Internal reliability estimates of Test No. 3 and its subtests are rather high for most subsamples, further suggesting the possibility of a valid measure of anaphoric resolution.

Tasks 3, 4, and 5

These tasks, embodied in Test No. 4, are intended to test sentence comprehension within a continuous text. It will be recalled that Task 3 requires the S to identify the sentence(s) in the text that provide the basis for a "true" or "false" response to the item statement. The judgment of such truth value constitutes Task 4.

To minimize the effects of random responses which may have yielded spuriously high scores on Task 4, Ss were presented with an additional response option for each item -- "cannot say". Of the twenty-three items, six were keyed "cannot say". In these instances, no transformation of the item sentence is contained in the text.

Given the nature of Task 1 (recognition of grammatical implication) and Task 2 (identification of anaphoric referents), it would seem reasonable that these two tasks might be better predictors of the criterion Task 5 and its components, Tasks 3 and 4, when only those items involving the variables defined for Tasks 1 and 2 are taken into account. An increased proportion of variance in Task 5 scores explained by Tasks 1 and 2 may also be expected.

To verify these propositions, zero order correlations were established and multiple regresssion analyses were conducted based on the group of seventeen items keyed either "true" or "false" which remained after the six items keyed "cannot say" were removed from the statistical analysis.

	Task 3	Task 4	Task 5
Task 1	.758	.666	.750
	(.717)	(.663)	(.706)
Task 2	.737	.644	.754
	(.711)	(.665)	(.720)

Table 29a: Correlation coefficients for 17 true/false items and (all items) for Native Speakers completing all tasks (N = 39, $p < .01$)

	Task 3	Task 4	Task 5
Task 1	.334	.452	.454
	(.275)	(.417)	(.390)
Task 2	.404	.417	.500
	(.380)	(.598)	(.625)

Table 29b: Correlation coefficients for 17 true/false items and (all items) for E.S.L.(A) Students completing all tasks (N = 44, $p < .01$)

	Task 3	Task 4	Task 5
Task 1	.427	.520	.401
	(.436)	(.540)	(.451)
Task 2	.359	.296	.226
	(.255)	(.201)	(.189)

Table 29c: Correlation coefficients for 17 true/false items and (all items) for E.S.L.(B) Students completing all tasks (N = 18, $p < .01$)

A comparison of these revised analyses with the original examination of the data which included all twenty-three items (Tables 29a, b, c) indicates that performance on Tasks 1 and 2 definitely tends to be a better predictor of indication of textual locus (Task 3) on a test that contains only items whose truth value can be determined directly from the text than to a test that contains at least some items whose truth value cannot be determined from the accompanying passage. This observation lends further support to the psychological reality of certain aspects of the model for comprehending sentences in a text as proposed earlier in this report (q.v., Appendix "A").

The effect of removing the indeterminate "cannot say" items from Test No. 4, however, is erratic with respect to predicting judgment of truth value, indicating that Task 4 is not as systematically related to the linguistic variables which define Tasks 1 and 2 as is Task 3.

A final consideration is the effect the removal of the indeterminate items from Test No. 4 has upon the amount of variance Tasks 1 and 2 contribute to the criterion measure. For native speakers, whose performance on Task 5 can be accounted to a greater extent by Tasks 1 and 2, the expected increase in variance explained by said tasks is realized (Table 30). In the case of E.S.L. students, for whom it has already been demonstrated that Tasks 1 and 2 contribute notably less variance to scores on the full Task 5, it can be seen that even less variance is explained by the two tasks on the abbreviated and more rigorously defined version of the criterion measure. This may well be a further indication that while native speakers utilize their proficiency in recognizing grammatical implication and anaphoric relations in responding to items in the format selected for Test No. 4, E.S.L. students rely less upon such capabilities, employing other

comprehension strategies to attain comparable scores on the criterion Task 5.

N.S.	E.S.L.(A)	E.S.L.(B)
70.4%*	30.9%*	16.2%*
(63.3%)*	(40.1%)*	(20.4%)*

Table 30: Combined variance contributed by Tasks 1 and 2 to criterion Task 5 scores comprising 17 true/false items and (all items) by Language group (*$p < .01$, ** not sig.)

Chapter VII

SUMMARY AND CONCLUSIONS

The major findings of the present study may be summarized:

1. There is a trend toward augmented performance on all tasks as a function of higher grade interval. (The literature reviewed in Chapter II cites evidence for developmental trends in syntactic comprehension.)

2. Native speakers as a whole tend to significantly outperform E.S.L. students only on Task 1 (recognition of grammatical implication). Analysis by grade intervals indicates that significant differences in Task 1 scores occur only at the secondary intervals.

3. No one subsample, even at the highest grade interval, demonstrates complete mastery of all transformations and anaphoric types tested. (Earlier investigations have discovered that school aged children are unable to comprehend many syntactic structures.)

4. No significant difference occurs at any grade interval between native speakers and E.S.L. students on the criterion measure of text comprehension or its component tasks.

5. The rank order of difficulty of the grammatical implication and anaphora subtasks differs for each subsample.

The performance of the multiethnic E.S.L. students more closely approximates the pattern of difficulty experienced by native speakers than does that of the Chinese E.S.L. learners, suggesting that native language may influence the order of syntactic difficulty for the second language learner.

The rank order of difficulty varies across grade intervals moreso for native speakers than for E.S.L. students. The rankings for elementary pupils across language groups is more consistent than for students at the secondary intervals.
(These findings all confirm the established principle that some syntactic patterns are easier to comprehend than others although one population may differ from another in this respect.)

6. In some instances, rankings appear to be affected by transformational simplicity. (Several researchers have claimed syntactic complexity to be a crucial factor in reading comprehension and readability.)

7. Regarding the components of text comprehension, native speakers find the task of judging the truth value of item statements to be significantly easier than identifying their locus in the text. For E.S.L. students these two tasks appear to be equally difficult. All language groups incur significantly lower mean scores on the criterion measure than on either of its component tasks.

8. The measures of syntactic comprehension are better predictors of the criterion for text comprehension and its component tasks for native speakers than for E.S.L. students. (The literature review and other discussion refer to a number of previous studies which have shown that syntactic comprehension contributes to more generalized measures of reading comprehension to varying degrees depending on the sample population.)

9. Indication of textual locus accounts for a greater proportion of the variance in criterion scores than does judgment of truth value in the

case of native speakers and the multiethnic E.S.L. students. The converse prevails for Chinese speakers.

Foremost among the present findings, however, is the evidence that native speakers and second language learners differ in the strategies used to attain equal proficiency on each of the component tasks of text comprehension. Specific differences can be related to a group's demonstrated linguistic knowledge. Those who possess one or another skill of syntactic comprehension draw upon that resource moreso than those who lack that skill. Hence, more than twice as much variance in the text comprehension criterion scores of native speakers can be accounted for by reference to Task 1 performance than can be explained for either of the E.S.L. samples. Conversely, more of the variance in the Chinese speaking E.S.L. group's criterion scores is explained by its performance on Task 2 which exceeds that of either of the other language groups. Results of this nature serve to further establish the principle of variations in the psycholinguistic processes through which the content of discourse is comprehended.

Directions for further research

Because the present study is of an exploratory nature, selection of transformational equivalents and anaphoric forms, while not entirely random, was, to some extent, eclectic. In reviewing other research instruments, comment was made on the degree to which each appeared to be motivated by a system of identifying and categorizing syntactic structures. The paucity of significant differences between native and non-native speakers among the present subtasks suggests that an extended investigation might be

facilitated by first attempting to build a comprehensive taxonomy of appropriate syntactic relations within a hierarchical framework initially conceived upon conceptual notions of transformational complexity. Through applied research, the proposed schema can be revised in such a manner as to extend beyond a mere categorized inventory of English syntactic patterns to derive an empirically developed schedule of acquistion difficulty. Once a sequence is clearly established for native speakers, it becomes practical to commence a systematic search for syntactic comprehension tasks which differentiate students at advanced stages of learning English as a second language from those of similar grade placement who possess native linguistic competence. The latter pursuit may be useful on two counts: (1) it could lead to more valid tests of English language competence than are presently available, and (2) having defined the deficiencies in syntactic understanding of the second language student, a basis might be provided for the design of compensatory instruction.

The broader problem remains, however, of establishing the relationship between syntactic tasks in isolation and those in a larger context. The present study is based upon a specific model for sentence comprehension within a text which demands that each test item have a syntactic relation to one or more sentences in the accompanying passage. It has been argued here and elsewhere that an adequate test of comprehension must be able to demonstrate which aspects of language competence are being tested by each item. Studies of passage dependency strongly suggest that conventional instruments, even those for which extensive norms have been developed, fail to meet this criterion. Therefore any attempt to correlate measures of syntactic comprehension designed to reflect understanding of selected aspects of language competence with what, by comparison, are loosely constructed inventories of

"reading comprehension skills" is not likely to reveal the full significance of one's knowledge of syntax in comprehending continuous discourse. Research linking syntactic comprehension as examined in this study with performance on popular standardized tests is, in the opinion of this investigator and others, predicated on faulty testing practices and therefore not worth the effort of careful and detailed study. It may be of some interest, however, to determine how tasks of grammatical implication and anaphoric resolution relate to a cloze test appropriate to the student's level of language competence or, in the case of the second language learner, a target level of competence.

On the basis of the present results, it would seem to be a productive course, from the standpoint of assaying the importance of syntactic knowledge, to further explore the relative contribution of locus indication and truth value judgment to the criterion measure of text comprehension submitted by this study. Research employing larger samples representative of a number of populations, particularly those considered to be below average in language comprehension as measured by conventional means, and using carefully constructed items based on a wide variety of syntactic variables may better evaluate the potential of the locus identification component and clarify a still ambiguous relationship.

Recommendations for testing

A number of concerns arise from the present study. The need for a comprehensive taxonomy of transformations and anaphoric forms has been mentioned in connection with directing motivated research. Testing that is motivated will also seek to select from an established taxonomy those grammatical implication relations appropriate to the requirements of a particular

assessment, i.e., level of syntactic complexity, formality of usage, subject of discourse, etc. Once a taxonomy becomes familiar to those engaged in the evaluation of language learning, it becomes possible for a publisher to describe in meaningful linguistic terms what areas of the syntactic system of English are examined by specific tests thereby enabling an evaluator to choose the instrument most appropriate to the situation and interpret the results in terms of specific deficiencies in the examinee's language competence.

In constructing tests intended to measure syntactic comprehension, two cautions must be kept in mind. First is a recognition of the interaction of syntax and semantics. Typical of recent experimental investigations in psycholinguistics is that of Herriot (1969) which found that the time required to interpret a syntactic relation (i.e., identify the subject of a clause) is affected by the absence of semantic clues. Similarly, O'Donnell (1976) has commented on the inappropriateness of using nonsense vocabulary in an attempt to gain a 'pure' measure of syntactic comprehension. It would seem that, since syntactic comprehension cannot occur unless one is able to recover the deep structure of a sentence, such recovery is halted by the inability to readily assign a lexical meaning to unfamiliar formatives. One is therefore well advised to devise only item statements that bear plausible propositions. Care was taken in the preparation of items for the present instrument to avoid complications introduced by semantic factors.

The second consideration is the importance of test format as an intervening variable. Examples of conflicting results have been cited between multiple choice and constructed responses in connection with investigations into the comparative difficulty of anaphora forms (Bormuth et al., 1970 vs. Lesgold, 1973) and between multiple choice and cloze exercises in

evaluating the effect of transformational simplicity on comprehension (Peltz, 1974). This suggests the need for caution in comparing performances on tests using different response formats. Admittedly such a problem arises in interpreting certain results of the present study. While comparisons across subsamples on the tasks of syntactic comprehension may be valid, it is debatable whether the cluster of low rankings for the anaphora subtasks is a reflection of actual comparative difficulty or a consequence of the variation between a multiple choice format in Task 1 and constructed responses in Task 2.[5]

Toward improving present practice, the task of indicating a textual locus for an item statement presents a new approach to comprehension testing, one that is based on observable linguistic features and, on the basis of the present evidence, one that is as valid as the conventional practice of asking the reader to make a judgment as to the truth value of a statement. While the need remains to further validate this task through additional research and analysis, its adoption by test constructors and classroom teachers enables writing items to test recognition of possible grammatical implication relations between any two sets of one or more sentences. All that is required is to choose a sentence, or sentences, in a passage on which to perform a selected transformation or anaphorization. The sentence resulting from the operation becomes the item statement. This approach eliminates the need to devise plausible false or indeterminate staements to balance the distribution of keyed responses. If the arrangement for numbering sentences in the passage is followed as in Test No. 4 of

---

5. Correlated t-tests between Task 1 and Task 2 conducted for each language group produced the following values: N.S.: t value = 3.96, $p < .001$; E.S.L.(A): t value = 2.42, $p < .025$; E.S.L.(B): t value = 4.53, $p < .001$.

the present instrument, items can be keyed for rapid scoring. With practice and caution in assuring that the keyed answers include all possible correct alternatives for the associated transformations of a particular item statement, the language teacher will acquire a measure that is likely to be both a valid indicator of achievement in English language learning and diagnostic of specific instructional needs in order to lead the student to native-like competence.

Implications for instruction

The present study is designed to identify some specific differences in language competence among instructional populations and to determine the extent to which the particular aspects of competence tested predict the ability to locate in a passage sentences related to a test item and to judge the truth value of the item statement. There is no intention here to prescribe programme objectives or teaching methods.

In light of the present results, the second language teacher must critically assay the importance of focusing attention upon transformational equivalents and anaphoric relations to instruction in discourse comprehension. E.S.L. teachers participating in this study apparently prepare their students through present instructional methods to perform tasks of text comprehension as adequately as native speaker peers. These tasks include both familiar objectives (Task 4) and unconventional demands (Task 3).

Examining a group of Grade 12 native speakers, O'Donnell concludes that "the correlation between awareness of structural relationships of words in sentences and ability in reading comprehension .44 ...is not sufficiently high to give conclusive evidence to support the teaching of linguistic structure as a major means of developing reading comprehension."(1963:316)

If one applies this argument to developing skill in deep structure recovery, the present data would suggest that such instruction would be appropriate only for native speakers.

Before accepting such a conclusion, however, one ought to consider the reports of some classroom applications. Hughes (1975) claims success in raising comprehension levels of Grade 7 students, particularly lower and middle ability groups, through practice in sentence combining. Implementing a curriculum to enhance comprehension of syntactic structures, Stedman (1971) was able to show significant gains in a cloze performance for Grade 4 black students who may resemble some second language learners in their unfamiliarity with syntactic forms prevalent in standard written English. In other attempts, O'Donnell and King (1974) were unable to improve the reading comprehension of Grade 7 students ranking below the 20th percentile on a standardized test through focused instruction in deep structure recovery. They do concede, however, that "DSRT Simons, 1970 abilities significantly influence reading comprehension test abilities" (p. 337). It is suggested that this sort of effort might be more productive with less debilitated students who show more positive attitudes toward a programme of remediation. Indeed, in a subsequent study, O'Donnell and Smith (1975) by designing a set of programmed exercises succeeded in increasing awareness of syntactic structure as measured by the PAST (O'Donnell, 1973) for some, but not all, of a group of Grade 9 students within four weeks of supplementary instruction.

While it is most probable that instructional content and delivery are crucial intervening variables in all programme studies, some evidence exists to support the contention that deep structure recovery skills as manifest in recognizing grammatical implication relations can influence

text comprehension and that, for at least some populations, such skills can be successfully taught.

If on this basis the E.S.L. teacher chooses to direct his students' attention to recognizing transformational equivalents and anaphoric relations, either because any such skills that identify native competence are worthy second language learning objectives in their own right or for what these abilities may contribute to more generalized measures of comprehension, he is well advised to be wary of using a model for testing as a model for instruction. Tests such as the DSRT, PAST, and the first two tasks of the present instrument are designed to sample from a large inventory the syntactic comprehension of selected forms by examinees at a given point in time. Referring to these tests, one cannot expect to list all relations that should be included in a programme; hence, the need to organize an E.S.L. curriculum upon a comprehensive instructional taxonomy. Further, it does not necessarily follow that practice on items adapted from the tests will over a period of time produce the linguistic understandings sought. Rather, the practicioner's task is to develop effective techniques and materials to provide lesson content, preferably in a context of graded texts and exercises which, unlike present publications, are ordered in their selection of clearly designated syntactic comprehension tasks.

BIBLIOGRAPHY

Baines, Helen. An assessment and comparison of syntactic complexity and word associations of good and poor readers in Grades 4, 8, and 12. 1975. ERIC microfiche ED 119 156

Bormuth, John R. Readability -- a new approach. Reading Research Quarterly, 1966, 1, 79 - 132.

---------- An operational definition of comprehension instruction. In Goodman, Kenneth S. and James T. Fleming (eds.) Psycholinguistics and the teaching of reading. Newark, Del., International Reading Association, 1969. pp. 48 - 60.

---------- On the theory of achievement test items. Chicago, University of Chicago Press, 1970

----------, John Manning, Julian Carr, and David Pearson. Children's comprehension of between- and within- sentence syntactic structures. Journal of Educational Psychology, 1970, 61, 5, 349 - 357.

Botel, Morton and Alvin Granowsky. A formula for measuring syntactic complexity: A directional effort. Elementary English, 1972, 49, 4, 513 - 516.

Carroll, John B. Comprehension by 3rd, 6th, and 9th graders of words having multiple grammatical functions. Final report. 1970. ERIC microfiche ED 048 311

Chomsky, Carol. The acquisition of syntax in children from 5 to 10. Cambridge, Mass., M. I. T. Press, 1969

Chomsky, Noam. Aspects of the theory of syntax. The Hague, Mouton, 1965

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

Cooper, Charles R. and Anthony R. Petrosky. A psycholinguistic view of the fluent reading process. Journal of Reading, 1976, 20, 3, 184 - 207.

Cox, Judy R. A study of the syntactic competence of adult beginning readers. 1976. ERIC microfiche ED 120 713

Cromer, W. The difference model: A new explanation for some reading difficulties. Journal of Educational Psychology, 1970, 61, 471 - 483.

Dale, E. and J. S. Chall. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11 - 20, 37 - 54.

Dunn, Olive J. Multiple comparisons using rank sums. Technometrics, 1964, 6, 3, 241 - 252.

Epstein, William. The influence of syntactic structure on learning. American Journal of Psychology, 1961, 74, 80 - 85.

Evans, Ronald V. The effect of transformational simplification on the reading comprehension of selected high school students. Journal of Reading Behavior, 1973, 5 4, 273 - 281.

Fagan, W. An investigation into the relationship between reading difficulty and the number and types of sentence transformations. Doctor's thesis. University of Alberta. 1969

Ferguson, George A. Statistical analysis in psychology and education. 2nd ed. New York, McGraw - Hill, 1976

Gibbon, Helen D. Reading and sentence elements. Elementary English Review, 1941, 18, 42 - 46.

Goodman, Kenneth S. and Carolyn L. Burke. A study of oral reading miscues that result in grammatical retransformations. 1969. ERIC microfiche ED 039 101

---------- Theoretically based studies of patterns of miscues in oral reading performance. Final report. 1973. ERIC microfiche ED 079 708

Granowsky, Alvin. Background for a new syntactic complexity formula. 1973. ERIC microfiche ED 083 566

Halliday, M. A. K. and R. Hasan. Cohesion in English. London, Longman's, 1976

Harris, Mary M. Second grade syntax attainment and reading achievement. 1975. ERIC microfiche ED 106 764

Hart, Margaret M. S. An assessment of the ability to manipulate syntactic structures as described by transformational grammar theory. 1971. ERIC microfiche ED 080 995

Herriot, P. The comprehension of active and passive sentences as a function of pragmatic expectations. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 166 - 169.

Ives, Sumner. Some notes on syntax and meaning. The Reading Teacher, 1964 18, 3, 179 - 183 + 222.

Kolers, P. A. Three stages in reading. In H. Levin and J. P. Williams (eds.) Basic studies in reading. New York, Basic Books, 1970

Kuntz, Mildred L. The relationship between written syntactic attainment and reading ability in 7th grade. 1975. ERIC microfiche ED 113 710

Lesgold, Alan M. Effects of pronouns on children's memory for sentences. 1972. ERIC microfiche ED 068 974

---------- Variability in children's comprehension of syntactic structures. 1973. ERIC microfiche ED 086 986

Marcus, Albert D. The development of a diagnostic test of syntactic meaning clues in reading. In R. Leibert (ed.) Diagnostic viewpoints in reading. Newark, Del., International Reading Association, 1971

Menzel, Peter. Linguistic bases of the theory of writing items for instruction stated in natural language. In J. R. Bormuth. On the theory of achievement test items. Chicago, University of Chicago Press, 1970

Miller, George A. Some preliminaries to psycholinguistics. American Psychologist, 1965, 20, 15 - 20.

---------- and S. Isard. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 217 - 228.

Mohan, Bernard A. Comprehension testing as semantics plus induction. Linguistics, 1973, No. 115, 93 - 104.

Morton, J. The effects of context upon speed of reading, eye movements, and the eye-voice span. Quarterly Journal of Experimental Psychology, 1964, 16, 340 - 354.

Nelson, Larry R. Guide to LERTAP use and interpretation. Dunedin, N. Z., University of Otago, Education Dept., 1974

Nie, Norman H., Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner, and Dale H. Bent. SPSS: Statistical Package for the Social Sciences. 2nd. ed. New York, McGraw - Hill, 1975

Oakan, R., M. Weiner, and W. Cromer. Identification, organization, and reading comprehension for good and poor readers. Journal of Educational Psychology, 1971, 62, 71 - 78.

O'Donnell, Roy C. A test of perception of syntactic alternatives. Studies in Language Education, Report No. 2. 1973. ERIC microfiche ED 077 025

---------- On linguistic structure and reading comprehension. 1976. ERIC microfiche ED 123 592

----------, William J. Griffin, and Raymond C. Norris. Syntax of kindergarten and elementary school children. Champaign, Ill., National Council of Teachers of English, 1967

---------- and F. J. King. An exploration of deep structure recovery and reading comprehension skills. Research in the Teaching of English, 1974, 8, 3, 327 - 338.

---------- and William L. Smith. Increasing 9th grade students' awareness of syntactic structures through direct instruction. Research in the Teaching of English, 1975, 9, 3, 257 - 262.

Pearson, P. David. The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. Reading Research Quarterly, 1974, 10, 2, 155 - 192.

Peltz, Fillmore K. The effect upon comprehension of repatterning based on students' writing patterns. Reading Research Quarterly, 1974, 9, 4, 603 - 621.

Reid, Jessie F. Children's comprehension of syntactic features found in some extension readers. In Jessie F. Reid (ed.), Reading: Problems and Practice. London, Ward Lock, 1972. pp. 394 - 403.

Richek, Margaret Ann. Effects of sentence complexity on the reading comprehension of syntactic structures. Journal of Educational Psychology, 1976, 68, 6, 800 - 806.

Rode, Sara S. Development of phrase and clause boundary reading in children. Reading Research Quarterly, 1974, 10, 1, 124 - 142.

Ruddell, Robert B. Language acquisition and the reading process. 1969. ERIC microfiche ED 033 819

Simons, Herbert D. Linguistic skills and reading comprehension. 1970. ERIC microfiche ED 047 927

Smith, William L. The effect of transformed syntactic structures on reading. 1970. ERIC microfiche ED 044 247

Spache, G. A new readability formula for primary grade reading materials. Elementary School Journal, 1953, 53, 410 - 413.

Stedman, Nathan A. III. The effect of a curriculum teaching syntactic embedding upon the reading comprehension of 4th grade students. 1971. ERIC microfiche ED 072 426

Steiner, R., M. Weiner, and W. Cromer. Comprehension training and identification of poor and good readers. Journal of Educational Psychology, 1971, 62, 506 - 513.

Stoodt, Barbara D. The relationship between understanding grammatical conjunctions and reading comprehension. 1970. ERIC microfiche ED 060 010

Takahashi, Barbara L. Comprehension of written syntactic structures by good readers and slow readers. 1975. ERIC microfiche ED 117 655

Tatham, S. M. Reading comprehension of materials written with select oral language patterns: A study at Grades 2 and 4. Reading Research Quarterly, 1970, 3, 406 - 425.

Tuinman, J. Jaap. Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 1973, 9, 2, 206 - 223.

van Metre, Patricia D. Syntactic characteristics of selected bilingual children. Claremont Reading Conference, 38th yearbook, 1974, 102 - 108.

Vogel, Susan A. Syntactic abilities in normal and dyslexic children. Baltimore, Md., University Park Press, 1975

Weber, R. M. First graders' use of grammatical context in reading. In H. Levin and J. P. Williams (eds.) Basic studies in reading. New York, Basic Books, 1970

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

Wisher, Robert A. The effects of syntactic expectations during reading. Journal of Educational Psychology, 1976, 68, 5, 597 - 602

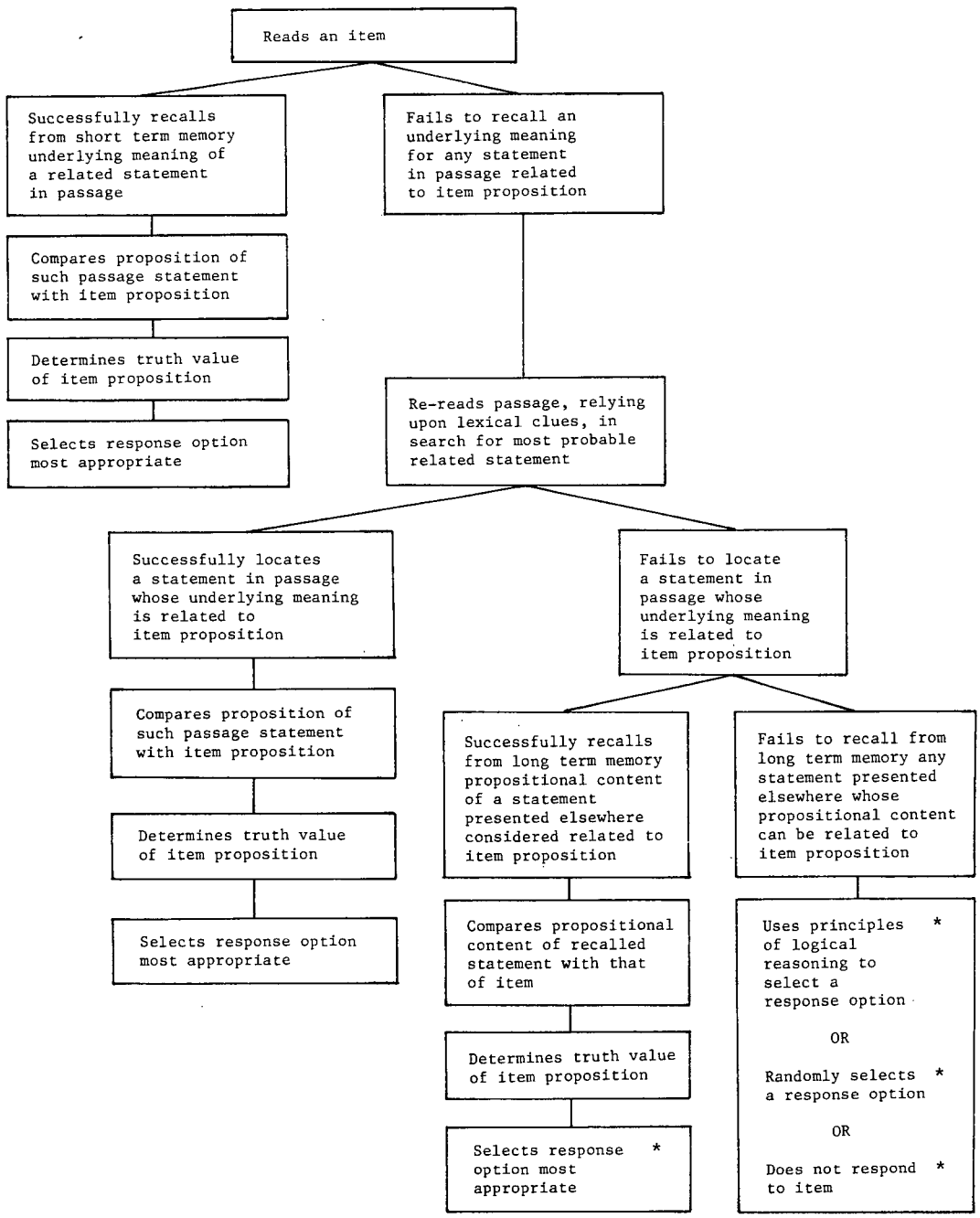
Appendix "A"

MODEL FOR SENTENCE COMPREHENSION WITHIN A WRITTEN TEXT

Test format: items in statement form, followed by three response options: true, false, cannot say

Assumption: propositional content of all item statements keyed "true" or "false" can be demonstrated to be derived from a specific locus in an accompanying passage through a series of syntactic transformations or anaphoric relations

Having read the passage in its entirety, the examinee



* a terminal behaviour that does not demonstrate comprehension of the text at hand

Appendix "B"

THE RESEARCH INSTRUMENT

Item difficulty

	Elem	JrSec	SrSec
N.S.			
E.S.L.			

N.B. In all cases, the item difficulty index is computed on the basis of the total number of Ss responding to the particular item. Not all Ss responded to all items or to both scales of Test No. 4.

Test No. 1

Directions: Read both sentences.
If the two sentences mean the same thing, circle "yes".
If the two sentences DO NOT mean the same thing, circle "no".

Sample A: The boy hit the girl.
The girl was hit by the boy. (yes) no

Sample B: The boy looked at the big dog.
The big dog looked at the boy. yes (no)

Subtask A: Passivization

1. A book was given to the girl by the boy.
The boy gave a book to the girl. *yes no

.917 1.00 1.00
.938 1.00 1.00

10. The buses were cleaned by the men.
The men cleaned the buses. *yes no

.833 1.00 1.00
1.00 .967 .947

19. The cat will chase the dog.
The cat will be chased by the dog. yes *no

.500 .700 .952
.813 .733 .895

28. The fireman saw the dog.
The fireman was seen by the dog. yes *no

.833 1.00 1.00
.938 .933 1.00

Subtask B: Participle modifier

2. The man saw that his car was stolen.
   The man saw his car that was stolen. yes *no

   .750 .800 .952
   .313 .067 .368

11. The man saw his stolen car.
   The man saw that his car was stolen. yes *no

   .667 .800 1.00
   .313 .533 .684

20. The man saw his car stolen.
   The man saw his stolen car. yes *no

   .750 .900 .952
   .438 .433 .632

29. The man saw his car stolen.
   The man saw his car being stolen. *yes no

   .333 .900 .857
   .500 .467 .737

Subtask C: Wh-- fronting

3. What John saw was a box.
   John saw a box. *yes no

   1.00 1.00 1.00
   .875 .900 .895

12. What the boy would like is for the girl to leave.
   The girl would like the boy to leave. yes *no

   .917 1.00 1.00
   .938 .867 .947

21. The man taught the boy to use a hammer.
What the man taught the boy was to use a hammer. *yes no

.667 .900 .952
.875 .933 .947

30. The boy wants the girl to find the ball.
What the girl wants is for the boy to find the ball. yes *no

.917 .900 .905
.813 .767 .947

Subtask D: Relativization contrasted with clausal conjunction

4. The woman called the policeman and he came down the hall.
The woman the policeman called came down the hall. yes *no

1.00 .700 .762
.938 .967 1.00

13. The girl hit the boy and he fell down.
The boy the girl hit fell down. *yes no

.167 .700 .810
.125 .267 .316

22. Helen drew a picture of a clown and went home.
Helen went home and drew a picture of a clown. yes *no

.833 1.00 1.00
.813 .900 1.00

31. Betty is happy and she likes her new school.
Betty likes her new school and she is happy. *yes no

.917 1.00 1.00
1.00 1.00 1.00

Subtask E: Relativization by pronoun deletion

5. The boy the girl hit fell down.
   The girl whom the boy hit fell down. yes *no

   .500 .700 .714
   .563 .567 .895

14. The horse which was brown stood behind the cow which was black.
    The brown horse stood behind the black cow. *yes no

    .917 .900 1.00
    .875 1.00 .947

23. The old tiger chased the young lion.
    The tiger which was young chased the lion which was old. yes *no

    .917 1.00 1.00
    .938 .900 .842

32. The man whom the teacher saw ran down the steps.
    The man the teacher saw ran down the steps. *yes no

    .833 .800 1.00
    .438 .500 .579

Subtask F: Double transformation (Relativization + Passivization)

6. The fat boy kicked the tin girl.
   The girl who is thin was kicked by the boy who is fat. *yes no

   1.00 1.00 1.00
   1.00 .967 1.00

15. The fat boy kicked the thin girl.
    The boy who is fat was kicked by the girl who is thin. yes *no

    .917 .500 1.00
    .938 .867 .789

24. The fat boy kicked the thin girl.
The girl who is fat was kicked by the boy who is thin. yes *no

.917 1.00 .952
.938 .967 1.00

33. The fat boy kicked the thin girl.
The boy who is thin was kicked by the girl who is fat. yes *no

.917 1.00 1.00
1.00 1.00 1.00

Subtask G: Ask (query) contrasted with Tell

7. The boy asked the girl what to do.
The boy asked the girl what she should do. yes *no

.667 1.00 1.00
.813 .700 .789

16. The woman told the man what to do.
The woman told the man what she should do. yes *no

.833 .900 1.00
.750 .700 .842

25. The woman asked the man what to do.
The woman asked the man what she should do. *yes no

.667 .800 .762
.813 .633 .737

34. The boy told the girl what to do.
The boy told the girl what she should do. *yes no

.750 .800 .810
.625 .767 .842

Subtask H: Easy to see

8. To see the girl is easy.
   The girl is easy to see. *yes no

   .917 1.00 .857
   1.00 .800 .842

17. It is easy to see the girl.
   The girl is easy to see. *yes no

   1.00 1.00 1.00
   .875 .767 1.00

26. The girl is easy to see.
   The girl sees easily. yes *no

   .917 1.00 1.00
   .563 .500 .684

35. The girl is easy to see.
   The girl is easily seen. *yes no

   1.00 1.00 .952
   .750 .400 .789

Subtask J: Promise contrasted with Tell

9. Mary told Jack to come here today.
   Mary told Jack that he should come here today. *yes no

   .917 .800 .762
   .750 .867 .895

18. Mary promised Jack to come here today.
   Mary promised Jack that he would come here today. yes *no

   .500 .900 .857
   .563 .600 .789

27. Mary told Jack to come here today.
Mary told Jack that she should come here today. yes *no

.917 .900 .952
.750 .900 .789

36. Mary promised Jack to come here today.
Mary promised Jack that she would come here today. *yes no

.417 .700 .762
.563 .667 .842

Test No. 2

Directions: Read the first sentence carefully. Then read the second sentence and decide if it is true or false.

Sample A: if: Ann and Helen walk to school together.
does it mean: Ann and Helen walk to school at the same time. (yes) no

Sample B: if: Mother said, "You must come home early.
does it mean: Mother must come home early. yes (no)

Subtask K: Indirect speech

1. if: Kathy said to her brother, "I want your skates."
does it mean: Kathy told her brother that he wanted her skates. yes *no

.833 1.00 .952
.750 .867 .789

2. if: Jack said to his sister, "Throw your ball to me."
does it mean: Jack told his sister to throw her ball to him. *yes no

.833 1.00 .952
.938 1.00 1.00

3. if: Susan said to Billy, "Do not come in."
does it mean: Susan told Billy that he should not come in. *yes no

1.00 .900 1.00
.875 .900 .895

4. if: Peter's brother said to him, "Do not talk in church."
does it mean: Peter's brother did not tell him to talk in church. yes *no

.750 .800 .810
.750 .867 .789

Subtask L: Ask (request) contrasted with Tell

5. if: Mike asks Bob to go first.
does it mean: Mike wants to go first. yes *no

.917 .900 .905
.875 .767 .947

6. if: Mike tells Bob to go first.
does it mean: Mike wants to go first. yes *no

1.00 1.00 1.00
.875 .833 .947

7. if: Mike asks Bob to go first.
does it mean: Mike wants Bob to go first. *yes no

.917 .900 .905
.938 .767 .947

8. if: Mike tells Bob to go first.
does it mean: Mike wants Bob to go first. *yes no

.917 1.00 .905
1.00 .733 .895

Subtask M: Pseudoimperatives

9. if I say: Sit down and I will scream.
does it mean: I will scream if you don't sit down. yes *no

.833 1.00 .952
.875 .933 .842

10. if I say: Telephone me on Friday or I will not come.
does it mean: I will not come if you telephone me on Friday. yes *no

1.00 .900 1.00
1.00 .633 .895

11. if I say: Turn on the radio or I will leave.
does it mean: If you don't turn on the radio, I will leave. *yes no

.833 .800 .952
.688 .667 .947

12. if I say: Come here and I will tell you.
does it mean: If you come here, I will tell you. *yes no

.917 1.00 .952
.938 1.00 .947

Subtask N: Agentless passivization

(N. B. Only item difficulties for the entire sample are reported here as this subtask was excluded from further analyses.)

13. if: The meat was eaten.
does it mean: Someone ate the meat. yes *no

.083

14. if: The money was stolen.
does it mean: Someone stole the money. *yes no

.750

15. if: The ball was thrown across the room.
does it mean: Someone threw the ball across the room. *yes no

.917

16. if: The chicken was killed.

does it mean: Someone killed the chicken. yes *no

.167

Test No. 3

Directions: (given orally)

Sample: I had an apple for lunch. It was good.

It apple

Subtask P: Pronominal reference

The dance of the bees tells which way to fly. It tells how far to go.

1. It (dance)

.333 .900 .952
.500 .433 .737

They picked up the bats and balls and put them away.

2. them (bats and balls)

.833 1.00 1.00
.938 .900 .947

Joe picked up the bat. He is a good hitter.

6. He (Joe)

.417 .900 .952
.813 .967 .737

A farmer knows that a hen will not lay an egg if it is shot.

11. It _____ (hen) _____

.333 .800 1.00
.813 .900 .842

Bees take the nectar that they like and make it into honey.

14. they ___ (bees) _____

.417 .800 .952
.875 .900 .842

15. it ___ (nectar) _____

.500 .800 .952
.750 .867 .789

To make butter, the cream is taken from the milk and set aside to sour. It is then churned until bits of fat come together.

16. It ___ (cream) _____

.333 .500 .952
.250 .300 .526

Subtask R: Pronominal reference (Minimal Distance Principle)

Jack said, "When he was six years old, Jim learned how to read."

4. he _____ (Jim) _____

.000 .600 .762
.375 .600 .158

Mary knew that Anne wanted her to pick up the toys.

8. her (Mary)

.417 .900 .952
.688 .900 .737

Helen told her mother that she was tired.

10. she (Helen)

.417 .800 1.00
.813 .933 .684

When Tom found out that Mike won the race, he was very happy.

13. he (Tom)

.333 .900 1.00
.813 .867 .737

Peter asked his father if he was hungry.

20. he (father)

.417 .900 .952
.438 .600 .789

Subtask S: Nominal substitution

Some dogs have collars with bells. Others do not.

3. Others (dogs)

.417 .900 .857
.563 .933 .684

Although the season for cherry blossoms is a short one, people can eat the fruit of the cherry tree all summer long.

12. one (season)

.333 .900 1.00

.563 .567 .579

For Chinese New Year, the windows which are made of thick rice paper are torn down and new ones are put up.

17. ones (windows)

.250 .800 .952

.188 .433 .421

In many cities there are buildings made of wood and others made of stone. Some are very old.

18. others (buildings)

.417 .800 1.00

.750 .800 .632

19. Some (buildings)

.417 .800 .857

.375 .800 .684

Subtask T: Clausal substitution

The boys played ball very hard. This is what won the game.

5. This (playing very hard)

.250 .600 .762

.250 .467 .526

Jim might come and play. The team hopes so.

7. so (Jim comes and plays)

.167 .500 .714

.438 .633 .842

Bill hurt his hand. This worried the team.

9. This (Bill's hurting his hand)

.500 .600 .810

.813 .967 .947

Test No. 4

(The following passage was presented on a separate page which was was detached from the test booklet.)

(1) Long, long ago, people did not know how to build houses.
(2) They had to live in caves on the sides of hills.
(3) They could keep themselves dry and warm in there.
(4) These people hunted wild animals for food.
(5) The skins of the animals were used for clothes.
(6) The cave men did not know how to write but they could draw.
(7) So they told many stories in pictures.
(8) They drew their pictures on the stone walls of the caves.
(9) In the last few years, some of these caves have been found.
(10) The picture stories are still there.
(11) They show animals and people of those early times long ago.
(12) They tell us things about cave men that we never knew before.

Directions: Read the story on the short paper first.
Read each question sentence below.

In the parentheses (____) write the number of the sentence, or sentences, in the story that tells you the answer.

If the question sentence is true, circle "T".

If the question sentence is false, circle "F".

If none of the sentences in the story tell you the answer to the question, put ✕ in the parentheses (____) and circle "?".

Sample A: The cave men did not use the skins of animals. (__5__) T (F) ?

Sample B: They lived in caves on the sides of hills where they could keep dry and warm. (2+3) (T) F ?

Sample C: The cave men built fires in front of their homes. (X) T F (?)

1. Long, long ago, no one knew how to build houses. (1) (T) F ?

Task 3	1.00	1.00	.950
	1.00	.967	1.00
Task 4	.833	.800	1.00
	.938	.967	1.00
Task 5	.833	1.00	.950
	.938	.931	1.00

2. People kept warm by living in caves on the sides of hills. (2+3) (T) F ?

Task 3	.417	.857	.850
	.438	.793	.889
Task 4	.833	.800	.900
	.875	.900	1.00
Task 5	.417	.857	.850
	.563	.828	.889

3. These people hunted food for wild animals. (4) T (F) ?

Task 3	.750	.714	.850
	.813	1.00	.889
Task 4	.583	.800	.800
	.563	.667	.789
Task 5	.500	.714	.800
	.438	.690	.778

4. The picture stories about cave men tell us new things. (10+12) (T) F ?

Task 3	.083	.571	.800
	.688	.517	.500
Task 4	.500	.900	.900
	.625	.500	.579
Task 5	.083	.571	.750
	.563	.448	.444

5. The cave men could raw because they did not know how to write. (6) T (F) ?

Task 3	.833	.571	.800
	.875	.897	1.00
Task 4	.417	.500	.250
	.188	.167	.000
Task 5	.250	.143	.250
	.188	.172	.000

6. In early times, all animals were wild. (X) T F (?)

Task 3	.500	.429	.700
	1.00	.897	.833
Task 4	.500	.400	.650
	1.00	.900	.842
Task 5	.500	.429	.650
	1.00	.897	.833

7. The picture stories are there still. (10) (T) F ?

Task 3	.417	.714	.950
	1.00	.931	1.00

Task 4	.750	.600	1.00
	.938	.900	1.00
Task 5	.417	.714	.950
	.938	.897	1.00

8. These people fed the wild animals. (X) T F (?)

Task 3	.500	.429	.550
	.813	.655	.833
Task 4	.417	.400	.550
	.813	.633	.737
Task 5	.417	.429	.550
	.813	.655	.722

9. Without drawing, these people used writing to tell stories. (6~~7~~) T (F) ?

Task 3	.083	.000	.400
	.063	.172	.222
Task 4	.750	.600	.800
	.313	.633	.526
Task 5	.083	.000	.400
	.063	.138	.167

10. Long, long ago, everybody had to hunt for his own food. (X) T F (?)

Task 3	.333	.286	.350
	.688	.621	.722
Task 4	.333	.400	.350
	.688	.600	.579
Task 5	.333	.286	.350
	.688	.621	.611

11. The cave men wore the skins of wild animals. (5) (T) F ?

Task 3	.667	1.00	1.00
	.625	.793	.833
Task 4	.750	.900	1.00
	.563	.800	.789
Task 5	.667	1.00	1.00
	.625	.759	.778

12. They hunted wild animals to eat. (4) (T) F ?

Task 3	.583	1.00	1.00
	.813	.931	.944
Task 4	.667	.900	1.00
	.750	.867	.947
Task 5	.583	1.00	1.00
	.813	.862	.889

13. The picture stories which show animals and people of early times are no longer there. (10+11) T (F) ?

Task 3	.000	.000	.250
	.063	.138	.111
Task 4	.500	.700	.950
	.375	.567	.737
Task 5	.000	.143	.250
	.000	.069	.111

14. The picture stories about cave men give us new information. (10+12) (T) F ?

Task 3	.583	.857	.900
	.188	.517	.556

Task 4	.667	.800	.950
	.250	.533	.684
Task 5	.583	.857	.900
	.188	.414	.500

15. There were wild animals in the caves. (X) T F (?)

Task 3	.583	.714	.700
	1.00	.897	.778
Task 4	.583	.600	.700
	1.00	.900	.684
Task 5	.583	.714	.700
	1.00	.897	.722

16. Pictures were one of the means of telling stories. (7) (T) F ?

Task 3	.250	.429	.900
	.313	.448	.778
Task 4	.833	.800	.900
	.313	.733	.789
Task 5	.250	.429	.850
	.313	.448	.722

17. How to build houses was known long, long ago. (1) T (F) ?

Task 3	.500	.571	.900
	.688	.690	.722
Task 4	.583	.700	.800
	.688	.667	.684
Task 5	.417	.571	.800
	.625	.655	.611

18. These people hunted wild animals for food. (4) (T) F ?

Task 3	.750	1.00	1.00
	.938	.966	.944
Task 4	.833	.900	1.00
	1.00	.967	.895
Task 5	.750	1.00	1.00
	.938	.966	.889

19. People sometimes drew pictures on the skins of animals. (X) T F (?)

Task 3	.667	.286	.600
	.688	.448	.722
Task 4	.667	.300	.600
	.688	.467	.632
Task 5	.667	.286	.550
	.688	.448	.611

20. Although they did not know how to write, the cave men could draw. (6) (T) F ?

Task 3	.583	1.00	.950
	.813	.931	1.00
Task 4	.917	1.00	1.00
	.750	.900	1.00
Task 5	.583	1.00	.950
	.750	.862	1.00

21. Hunted by these people, the wild animals were eaten. (4) (T) F ?

Task 3	.583	.714	.900
	.375	.724	.833
Task 4	.667	.700	.850
	.313	.600	.842

Task 5	.583	.714	.850
	.313	.621	.833

22. We can learn new things by looking at the picture stories about cave men. (10+12) (T) F ?

Task 3	.583	.714	.850
	.375	.621	.833
Task 4	.667	.700	.950
	.563	.600	.789
Task 5	.583	.714	.850
	.438	.552	.833

23. Some of the picture stories will never be found. (X) T F (?)

Task 3	.750	.857	.850
	.625	.655	.722
Task 4	.750	.800	.750
	.625	.633	.579
Task 5	.750	.857	.750
	.625	.655	.611