DOCUMENT RESUME

ED 052 914 24 RE 003 765

AUTHOR	Auerbach, Irma-Theresa
TITLE	An Analysis of Reading Comprehension Tests. Final Report.
INSTITUTION	Harvard Univ., Cambridge, Mass. Graduate School of Education.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-0-A-074
PUB DATE	Jun 71
GRANT	OEG-1-71-0001
NOTE	214p.; Doctoral Dissertation
EDRS PRICE	EDRS Price MF-$0.65 HC-$9.87
DESCRIPTORS	Achievement Tests, Doctoral Theses, Item Analysis, Measurement, Measurement Instruments, *Readability, *Reading Comprehension, *Reading Research, Reading Skills, *Reading Tests, Test Construction, *Test Validity

ABSTRACT

An investigation of reading comprehension tests was undertaken in this doctoral dissertation study to find out what the tests evaluate. The history of standardized reading tests and of efforts to define comprehension was traced. Standardized tests from three batteries (California Achievement Test, Gates-MacGinitie Reading Test, and Stanford Achievement Test) were analyzed at three grade levels (1 to 2, 4 to 6, 9 to 14). The first analysis, for readability, used the Dale-Chall and the Spache formulae, and the second analysis, for difficulty of test items, used specially designed rating scales. Tests differed on all readability counts and scores increased with higher grade levels, although not always at the same rates either across tests or within them. Test items tended to use different types of questions for similar kinds of materials. Questions fell generally into paraphrase and concept types. Frequency data were calculated for each of the readability and difficulty measures, and correlation coefficients were calculated for several characteristics of individual tests. While differences were found among the tests, they appeared to be testing abilities similar to those evaluated by intelligence and achievement tests. A model for constructing new and better defined reading comprehension tests, a bibliography, and tables are included. (MS)

FINAL REPORT

Project No. 0-A-074
Grant No. OEG-1-71-0001

AN ANALYSIS OF READING COMPREHENSION TESTS

Irma-Theresa Auerbach

Harvard College
Graduate School of Education

Cambridge, Mass. 02138

June 1971

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RE 003 765

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

ANALYSIS OF STANDARDIZED READING COMPREHENSION TESTS

Irma-Theresa Auerbach

A Thesis Presented to the Faculty of the
Graduate School of Education of Harvard University
in partial fulfillment of the requirements for
the Degree of Doctor of Education

1971

ACKNOWLEDGEMENTS

Appreciation is due Dr. William E. Kline of the California Test Bureau, Professor Walter MacGinitie of Teachers College, Columbia and the Test Division of Harcourt, Brace and World, Inc. for kindly providing the tests that were analyzed in this study and related information that I frequently needed.

I would like to take this opportunity to express my sincere gratitude to Professor Jeanne S. Chall for both wise tutelage and generous support during my studies at Harvard. Professor Douglas Porter exhibited unending patience and guidance throughout the development of this thesis, and Professor Israel Scheffler made many suggestions which greatly helped to focus and to clarify the ideas presented in this work.

Joshua Copel of the Roxbury Latin School contributed his independent study time to assist in this research. He has truly become a "Master of Tables."

My many friends not only offered encouragement but some, especially Dr. Sidney Kadish, took time out of their work to read and criticize this dissertation.

Most of all I wish to thank my mother whose many sacrifices have made my education and this study possible.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

Figure Page

THESIS ABSTRACT

Analysis of standardized reading comprehension tests

Reading comprehension skill builders and tests seem to reflect different concepts of comprehension. A review of the history and development of reading comprehension tests, and the research in comprehension revealed that no concise or agreed upon definition of reading comprehension exists.

Despite the lack of clear definition, tests of reading comprehension are frequently used in evaluating the performance of pupils and teachers as well as the effectiveness of instructional materials and methods.

This study was designed to investigate what reading comprehension tests evaluate, i.e. what pupils must do or know to perform well on selected standardized reading comprehension tests. Standardized reading comprehension tests were selected at three levels (Grade 1-2, 4-6, 9-14) and from three batteries (California Achievement Test, Gates-MacGinitie Reading Test and Stanford Achievement Test).

Two types of analyses were conducted. The first analysis was a study of the readability of reading comprehension test items. Two widely-used readability formulae were employed -- Dale-Chall (1948) and Spache (1953). The second analysis was a study of tasks required by reading comprehension test items. The measures of task were designed

for this study and included a rating scale for the reading selections, a rating scale for the questions and a rating scale for the choices. Distinct and differing characteristics emerged for both readability scores and task ratings among the three test levels and the three test batteries analyzed.

Tests were found to differ on essentially all readability counts, e.g. One test had an average selection length of over 400 words while another at the same test level but from a different test battery had an average of less than 60 words. However, complementary relationships seemed to exist, e.g. while one test had long reading selections and short questions, another test had short selections and long questions. Also, readability scores consistently increased with higher test level. For example, reading selections, questions and choices were usually longer and had more hard words at higher test levels.

The task analysis revealed that different test batteries contained somewhat similar types of reading selections, differed considerably on types of questions and had somewhat similar distractors. At lower test levels selections were generally about common incidents and people. At higher test levels, selections were more about academic subjects such as science or social studies. Test questions were of two major types: paraphrase and concept. Paraphrase questions included eight kinds of restatements of given information, e.g. contextual paraphrase, grammatical paraphrase. Concept questions included six categories and always applied when all the information was not given in the reading selection, e.g. probable concept, language concept, previous knowledge of science.

Whereas selected reading comprehension tests were found to differ in what they were testing, they appeared to be testing abilities similar to those evaluated by I.Q. tests and achievement tests in other school subjects.

The findings of the analyses suggested a model for new and better defined reading comprehension tests. Such tests would include "criteria" and descriptions of the following five features:

1. length, sentence length, and hard word ratio of reading selections, questions and choices
2. topics of reading selections
3. tasks necessary for supplying the correct answer to the question
4. types of distractors provided as alternate answers.

CHAPTER I

The Problem

Any child who fails to acquire the ability to read has been denied a right--a right as fundamental as the right to life, liberty, and the pursuit of happiness--the right to read....

It is inexcusable that in this day when man has achieved such giant steps in the development of his potential, when many of his accomplishments approach the miraculous, there still should be those who do not learn to read....

Therefore, as U.S. Commissioner of Education, I am herewith proclaiming my belief that we should immediately set for ourselves the goal of assuring that by the end of the 1970's the right to read shall be a reality for all-- that no one shall be leaving our schools without the skill and the desire necessary to read to the full limits of his capability.

James E. Allen, Jr.

Introduction

The ultimate goal of reading instruction is to develop reading comprehension (Chall, 1967, p. 307). Unfortunately, there is as yet no concise or agreed upon definition of what comprises reading comprehension. Consequently, no consistent means exist of either teaching or testing it.

This thesis illustrates the problem by demonstrating the existence of major differences among a sample of reading comprehension skill builders and among selected reading comprehension tests.

An attempt to find a more consistent definition of reading comprehension has been undertaken here by systematically analyzing standardized reading comprehension tests. These empirically constructed tests have long been the accepted criterion for establishing success or failure in reading comprehension. A clarification of what tests are actually testing should contribute to a clarification of what is currently meant by reading comprehension.

Current Trends in Teaching Reading Comprehension

An analysis of current trends in teaching reading comprehension will demonstrate some of the confusion that exists in defining the concept.

Typical materials used to teach comprehension are collections of reading selections (either in basal readers with accompanying workbooks, or booklets, or boxed packages called "reading laboratories").[1]

[1] A sample of widely used basal readers with workbooks, booklets and laboratories that teach comprehension are:

Title	Publisher and Date		Reading Grade Level
Basal Readers with Workbooks			
Basic Readers	Ginn	1969	Pre-primer - 6
Basic Reading	Lippincott	1964	Pre-primer - 8
Basic Reading Program	Harper & Row	1966	Pre-primer - 6
Macmillan Reading Program	Macmillan	1966	Pre-primer - 6
New Basic Readers	Allyn & Bacon	1968	Pre-primer - 6
Open Court Basic Readers	Open Court	1967	Pre-primer - 6
Sheldon Basic Reading Series	Allyn & Bacon	1968	Pre-primer - 8
Booklets			
Be a Better Reader	Prentice-Hall	1968	4-12
Macmillan Reading Spectrum	Macmillan	1964	4-6
McCall-Crabbs Standard Test Lessons in Reading	Teachers College	1961	2-12
New Practice Readers	McGraw Hill	1961	2-8
Read for Meaning	Lippincott	1955	4-12
Readers Digest Skill Builders	Readers Digest	1963	1-8
Reading Exercises	Teachers College	1965	2-6
Reading for Concepts	McGraw-Hill	1970	1-6
Specific Skill Series	Barnell Loft	1967	1-6
Laboratories			
Reading Attainments System	Grolier	1967	3,4 (easy reading intended for older pupils and adults)
Reading Laboratory	Science Research Associates	1961	2-7
Reading for Understanding	" " "	1958	5 - college

The selections are generally graded in difficulty by field-testing them on children of various grades, by asking expert opinion, or by one of the widely used readability formulae such as the Dale-Chall (1948), Flesch (1948), or Spache (1953). Comprehension questions follow the selections. These are usually multiple-choice or completion questions that ask the student to identify or relate the "main idea," "facts," or "inferences."

Table 1 summarizes some aspects of five randomly selected skill builders in booklet form. Generally the information in Table 1 was taken directly from teachers' manuals, although in some cases the "topics" and "questions" were not explicitly stated in the manual. "Topics" and "questions" were then established by reviewing the skill builders themselves.

The differences in the structure and content of these skill builders appear to reflect the differences in the authors' conceptions of reading comprehension. For example, in "purpose" (see Table 1) New Practice Readers proposes to develop seven "elements in comprehension," while Reading Exercises proposes to develop speed, general comprehension and three "specialized skills." In "questions," Standard Test Lessons in Reading has only multiple-choice questions, while Be A Better Reader has primarily open-ended questions (questions for which no answer choices are provided).

Four major related issues emerge from an analysis of the purposes of reading comprehension skill builders. First, skill builders use different instructional strategies for teaching comprehension.

Table 1

Comparison of Five Comprehension Skill Builder Booklets

	Standard Test Lessons In Reading (1950)	New Practice Readers (1962)	Reading Exercises (1965)	Be A Better Reader (1968)	Reading for Concepts (1970)
PURPOSE	...help pupil develop ...speed and...comprehension (p. 3)	...maximize development of seven elements (p. 1; see elements below)	...develop speed, general comprehension, and these three more specialized reading skills (p. 3; see skills below)	...skill development in reading in the content subjects (p. T4; for content subjects, see topics below)	...grow in reading experience while exploring a wide variety of ideas (p. 3)
LEVELS	2-12 (field trials; p. 3, 13)	2.5-6.8 (Spache and Dale-Chall readability formulae, p. 3)	2-6 (p. 4) expert opinion	4-high school (Dale-Chall, Flesch and SRA readability formulae, p. T4)	1.6-6.7 (Spache and Dale-Chall readability formulae, p. 5)
TOPICS	animals, plants, people, city life, farm life, games, etc. (p. 3)	social studies, nature, scientific achievements, character development, safety practices, and health (p. 5)	similar in character and length to newspapers, magazines, books of games cookbooks, encyclopedias, advertisements, editorials, etc. (p. 4)	social studies, science new mathematics, and literature (p. T4)	anthropology, art, biology, earth science, ecology economics, engineering, geography, history, mathematics, political science, sociology, space (p. 5)
STYLES	stories, fables, poems, informational selections, directions, etc. (p. 3)	factual articles, folk tales, adventure stories, (p. 6)	(see topics above)	stories, narrative accounts, instructions (p. T3-T15)	stories (p. 5)
TESTS	field-tested all skills (below) together	all skills (below) together	three skills (below) separately	separately for content subjects (topics above) and for skills (below)	field-tested skills (below) separately
QUESTIONS	multiple-choice	sentence completion, multiple-choice, true-false (p. 5)	multiple-choice, open-ended	open-ended, multiple-choice	sentence completion, multiple-choice
SKILLS, ELEMENTS, ABILITIES, ETC.[a]	stated fact inference main thought etc. (p. 13)	direct detail implied meaning whole meaning affirmative, negative, or unstated idea substantiating from text antecedents word meaning from context (p. 4)	details main idea following directions (p. 3-4)	direct, literal meaning interpretation passing personal judgment on quality, accuracy or value of text projecting thinking beyond text skills specific to content subjects (p. T16)	fact inference or implication main idea interpretation, modification substantiating from text antecedents, antonyms word meaning from context cause and effect concept (p. 6)

[a] Words and phrases such as "recalling, recognizing, understanding, finding, ability to, competence in," etc. were deleted from the manuals' description of skills (above) because they seemed to be used interchangeably and thus obfuscated rather than clarified the skill descriptions.

Some skill builders concentrate on exercising testing procedure. For example, Standard Test Lessons in Reading emphasizes increasing reading speed and the number of correct answers by providing practice in test-like exercises. Other skill builders seem to emphasize increasing the readers' background knowledge. For example, Reading for Concepts provides the reader with organized information about selected disciplines. The Standard Test Lessons in Reading approach suggests that comprehension is improved most by continuous practice in answering certain types of questions (similar to those on tests), while the latter approach suggests that comprehension is improved most by giving the student specific types of information or subject matter knowledge.

Second, some skill builders concentrate on general topics, e.g. Standard Test Lessons in Reading includes selections about animals, city life, plants, people, etc. in no apparent sequence or proportion, thus implying that comprehension "skills" are general to many types of reading matter. Other skill builders carefully differentiate among topics, e.g. Be a Better Reader presents exercises in 4 disciplines: social studies, science, new math, literature, and in so doing implies that comprehension "skills" are different for different subjects or disciplines.

Third, some skill builders isolate specific reading comprehension "skills," such as Reading Exercises, which presents special booklets and exercises for identifying "details," for finding the "main idea," and for "following directions." This suggests that comprehension is composed of numerous separate subskills. Other skill builders combine many skills, e.g. Standard Test Lessons in Reading mixes questions about "stated facts," "inferences" and "main thoughts" into one exercise and booklet in no apparent sequence or proportion. This approach suggests that comprehension is more of a general skill than a combination of clearly defined subskills.

Fourth, some materials have recently become available, usually for fourth grade and higher levels, that attempt to give instruction to the student in how to go about answering certain types of questions. For example, Be a Better Reader (1968, p. 4) instructs pupils to know "who the people in the story are...where they are, what they did, and what happened to them" from words directly stated in the reading selection. Such instruction is intended to help pupils understand facts.[1]

Most materials do not provide adequate measures to determine and treat types of reading difficulties.[2] If a pupil consistently answers questions incorrectly there is usually no suggestion given to either pupil or teacher other than continued exercise of the same kind. Indeed, the basic assumption underlying most instructional materials and methods is that comprehension can be induced by raising, or

[1] Some inadequacies of this technique of teaching reading comprehension are presented in Simons (1970), Chapter 1.

[2] Reading difficulties as demonstrated by errors on these and similar reading exercises may be due to a misunderstanding or misinterpretation of the selection or question (Thorndike, 1914). Errors may also be due to a deficiency in word recognition (Thorndike, 1915; Chall, 1958a). Since exercises are read silently and questions are answered independently the source of error usually remains undetermined.

lowering in the case of problem learners, the readability level of the reading matter presented.

The fundamental issue here is whether reading comprehension is an analyzable skill that can be reliably diagnosed and directly taught; or whether it is an unanalyzable skill that can merely be exercised in a general way.

The lack of agreement within the category of purpose of skill building exercises, as described above, also exists within the other categories listed in Table 1. The "skill" category of Table 1 contains only vague descriptions given in the skill builders' teachers' manuals. Language used to describe the skills in Table 1 for different skill builders may be identical, e.g., both Reading Exercises and Reading for Concepts list as one of their skills "main idea." Unfortunately, it is not clear that the corresponding tasks are, indeed, identical. In addition, the lists of skills seem to confuse instructional procedures and psychological processes with comprehension skills.[1] For instance, while "interpreting" may be a comprehension skill that can be developed through instruction and exercise, "inference" may be more of a psychological process that can not be readily modified. And, "finding the main idea," "locating details," and "following directions" generally seem to be instructional "sets" or procedures used by teachers or authors of skill builders to exercise comprehension. While exercise materials generally represent instructional strategies and procedures, their relationship to

[1] Simons, op. cit.

psychological processes and comprehension skills remains enigmatic.

It appears that materials designed to teach reading comprehension are inconsistent in their underlying hypotheses about the nature of comprehension and in the types of "lessons" designed to exercise it. In addition, these materials generally do not afford the teacher or pupil an analysis of strengths and weaknesses, nor do they usually provide instructional procedures other than the selection-question exercise. Consequently, most materials designed to teach reading comprehension actually are tests of comprehension arranged by successive levels of difficulty.

Current Trends in Testing Reading Comprehension

The issues and problems of instructional materials in comprehension also apply to tests of reading comprehension. Unfortunately, the problems involved in testing comprehension are even more serious since tests represent the criteria of competence in comprehension. In other words, comprehension is generally defined by what is tested on tests of reading comprehension.

The results of these tests have considerable educational and social significance. Reading comprehension test scores are used for accepting pupils into schools, putting pupils into special classes, grouping pupils within classes, determining promotion, acceleration or demotion, presenting academic awards, counseling for future education and in some cases setting teacher expectation. Furthermore, new educational materials as well as many government and industry-sponsored educational programs as well as teachers and methods are evaluated mainly on the results of these tests. Annually many millons of dollars from the educational budget are appropriated solely for the purchase and scoring of standardized tests.[1]

Standardized tests are almost universally used in schools to assess pupils' reading ability (Stevens, 1971). These tests are generally constructed on the basis of three assumptions. The assumptions are that older children and children with higher I.Q.s perform better; and that performance in comprehension follows the "normal distribution" model. The "normal distribution" model generally assumes that performance in reading comprehension is superior in 4% of the population of students at each grade or test level, above average in another 19% of the population, average in 54% of the population, below average in 19% of the population and poor in another 4% of the population (Kelley, et al, 1966, p. 10).

Thus, a large number of experimental test items are administered to many pupils at many grade levels. Either all or a sample of these pupils are also given I.Q. tests.[2] Test items that are found to discriminate empirically, for whatever reason, among older and younger

[1] The cost of the first state-wide standardized achievement testing program in Massachusetts, which included about 100,000 fourth graders, was $120,000.00 (Cohen, 1971, p. 1,5).

[2] For example, pupils tested for the California Achievement Tests were also given the California Test of Mental Maturity (California Test Bureau, 1957, p. 18-20); some pupils given the Gates-MacGinitie Reading Tests were also given the Lorge-Thorndike Intelligence Tests (Gates and MacGinitie, 1969, p. 1-2); and pupils given the Stanford Achievement Test were also given the Otis Quick-Scoring Mental Ability Test (Kelley, et al, 1966, p. 9).

pupils and among pupils with high and low I.Q.s, as well as items that have acceptable empirical difficulty scores are considered for inclusion in the final form of the tests. The empirical difficulty score of an item is the proportion of pupils in a given population that answered the item correctly. Items are included in tests so that the combination of difficulty scores forms a "normal distribution." For example, a relatively small proportion of items passed empirically, for whatever reason, by a small proportion of pupils is chosen. These items are considered difficult and probably make up about 23% of the items in the test. A relatively large proportion of items, 54%, is considered of average difficulty and includes items that were passed by approximately 70-80% of the population, and so on, until a "normal distribution" appears.

After the test items are chosen standardization procedure requires selecting large student samples representative of the national school population (Kelley, et al, 1966, p. 9). Representativeness is usually based on census data and includes such population characteristics as geographic distribution, community or school size, median family income, median number of years of schooling completed by those over 25 years of age, chronological age by grade, and mental ability of the group (California Test Bureau, 1957, p. 12; Gates and MacGinitie, 1965d, p. 2; Kelley, et al., 1966, p. 9-10.) Tests are uniformly administered to pupils in this population. Norms for test scores at given grade levels are calculated. Common scores are grade scores, percentiles and stanines (Kelley, et al, 1966, p. 10)[1]

Consequently, standardized test scores provide only a rank of pupil performance on a given task in relation to the standardization population. No matter how literate a population is, the lower 23% always scores below "grade level."[2] Furthermore, no one seems to know why one item is harder than another, what a pupil must do or know to answer the question and so on. Generally standardization procedures seem to receive a disproportionate amount of description and discussion in test manuals at the expense of more qualitative aspects of the tests.

Table 2 summarizes several "qualitative" aspects of three standardized reading comprehension tests. Information for Table 2 was taken from teachers' manuals and technical bulletins that accompanied the tests. Differences among tests are apparent.[3] Again, the differences in the structure and content of comprehension tests reflect the differences in their authors' conceptions of reading comprehension.

As noted on Table 2, purposes and skills for tests of comprehension seem even more vague than those of instructional materials (on Table 1) e.g., skill builders enumerated the skills they included in comprehension:

[1] Dr. Henry Dyer (1971, p. 19) has attacked these scores calling them "psychological and statistical monstrosities" because they are so easily and so frequently misinterpreted. Grade scores for example vary from test to test. One test may put a pupil's reading performance at 4th grade while another test will rate the pupil at 5th grade. Sometimes 2 or 3 wrong answers change a grade score by one year. Population samples were also criticized for often not being representative of the nation.

[2] Kelley, et al, (1966, p. 29) stated that the 1964 Stanford Achievement Test, for instance, yields "harder norms" than the 1940 or 1953 editions of the test. For example, in 1940, 4th grade level corresponded to a test score of 12. In 1964, 4th grade level corresponded to a test score of 18.

[3] For a penetrating analysis of how tests differ see Kerfoot (1965).

Table 2

Comparison of Three Standardized Reading Comprehension Tests

	California Achievement Test (1963)	Stanford Achievement Test (1964)	Gates-McGinitie Reading Test (1965)
	Interpretation Subtest	Paragraph Meaning Subtest	Comprehension Subtest
PURPOSE	...designed to reveal the pupil's comprehension of what he reads and to enable the teacher or counselor to make a diagnosis of specific difficulties (Tiegs and Clark, 1963a, p. 5)	...functional measure of pupil's ability to comprehend connected discourse (Kelley, et al, 1964a p. 5)	...measure the student's ability to read complete prose passages with understanding (Gates and MacGinitie, 1965b, p. 1)
LEVELS	1-14 (California Test Bureau, 1957, p.5)	1-12 (Kelley, et al,1964a, p. 3; Gardner, et al, 1965, p. 7)	1-12 (Gates and MacGinitie, 1965d, p. 1)
TOPICS	animals, industrial process (Tiegs and Clark, 1963b, p. 6) geography, economics, science, psychological tests, philosophy (Tiegs and Clark, 1963a, p. 6)	general reading material, life science, physical science, geography, history, literature, social science, fine arts (Kelley, et al, 1964a, p. 5)	science, social studies, sports, stories, etc.
STYLES	directions, stories (Tiegs and Clark, 1963c, p. 5)	paragraphs (Kelley, et al, 1964a, p. 5)	whole sentences and paragraphs (Gates and MacGinitie, 1965a, p. 1)
TESTS	normed (C. T. B., 1967, p. 11-18) all skills(below)together	normed (Kelley, et al, 1966, p. 9-16) all skills (below) together	normed (Gates and MacGinitie, 1965d, p. 2)
QUESTIONS	separate questions with either no answer choices or multiple choice answers (C. T. B., 1957, p. 5)	blanks in paragraph and separate questions; both with multiple-choice answers (Kelley, et al, 1964a, p. 5)	blanks in paragraph with multiple-choice answers (Gates and MacGinitie, 1965b, p. 1)
SKILLS	organization of topics interpretation of material directly stated fact inferences topic or central idea sequence of events (C. T. B., 1967, p. 35-37) follow directions (Tiegs and Clark, 1963c, p. 7)	...varying from extremely simple recognition to the making of inferences from what is stated in several sentences (Kelley, et al, 1964a, p. 5)	mark the picture that best illustrates the meaning of the passage or that answers a question in the passage (Gates and MacGinitie, 1965a, p. 1) decide which one of five completions best conforms to the meaning of the whole passage (Gates and MacGinitie, 1965b, p. 1)

Reading Exercises named "details", "main idea," and "following directions." Test-authors seemed less concerned with specificity. One test author merely stated that the test was evaluating "extremely simple recognition to the making of inferences (Kelley, et al, 1964a, p. 5)." Hence only very limited information about the authors' conception of reading comprehension could be gained from test descriptions or manuals. Chall (1967, p. 312) has noted that "standardized reading tests often mask some of the important outcomes of instruction because they measure a conglomerate of skills and abilities at the same time."

This confusion about reading comprehension is summed up by Kolers"

> We cannot yet describe accurately even what it is we are measuring when we measure "comprehension" in reading tests, or what we mean by "understanding," and we cannot yet say accurately what it is we mean by "meaning"....(Kolers, 1968, p. xxxiv)

Despite the lack of knowledge with regard to the teaching, testing and nature of reading comprehension, many individuals do read adequately to meet the needs of everyday life. Obviously, then, something in the educational experience of these individuals has been effective. Analysis of the skills of effective readers may contribute to an understanding of reading comprehension generally.[1]

Despite their limitations, standardized reading comprehension tests

[1] Smith (1971) presented a model of the reading comprehension process derived from an analysis of mature reading. The approach presented here differs in that "comprehension achievement" as demonstrated by empirically constructed standardized tests at the elementary, intermediate and advanced grade levels is analyzed.

offer a source of empirical data for understanding "reading comprehension." Norming procedures, including detailed analyses of item difficulty, produce an empirically valid progression of reading performance. An analysis of reading comprehension tests holds promise for revealing

1. the nature of the comprehension task
2. whether the task differs by grade level
3. whether the task differs by test battery
4. what determines difficulty of the task.

Summary

Because there seems to be no clear or consistent definition of reading comprehension, significant differences appear among materials designed to teach and test it. A study of comprehension tests will reveal what tasks are currently used as the criteria for comprehension.

The remaining chapters of this dissertation will consist of an analysis of the historical development, structure and content of selected standardized reading comprehension tests. The analysis includes investigation of language and performance factors in selections, questions and choices of tests.

Chapter II presents the development of standardized reading comprehension tests. Chapter III briefly introduces the objectives of this dissertation, the reading comprehension tests that were studied, and the analyses that were conducted. Chapter IV presents a comprehensive study of the readability of reading comprehensioa tests. Chapter V presents a comprehensive study of the tasks in reading comprehension tests. On the basis of these analyses a suggestion is made for new tests of comprehension in Chapter VI.

CHAPTER II

Development of Standardized Reading Comprehension Tests

Historical Foundations in Testing

Edward L. Thorndike initiated standardized testing in reading comprehension. In 1914 he published a major article on the topic in the Teachers College Record entitled "Measurement of Ability in Reading." He began by experimenting with different "degrees" or types of understanding:

> ...What degree of understanding we require in our test is of almost no consequence, but that we should define objectively the degree of understanding that we do require is of very great importance.... (Thorndike, 1914, p. 226)

Toward this end, he devised the "Visual Vocabulary Scale". This set of tests did "not measure ability to understand the meaning of these printed words in general, or, as they come in ordinary texts, or completely, but only to understand them well enough to classify them as required by the test. (Thorndike, 1914, p. 226)":

> Look at each word and write the letter F under every word that means a flower.
> Then look at each word again and write the letter A under every word that means an animal.
> Then look at each word again and write the letter N under every word that means a boy's name.
> ...
> ...
> 4. camel, samuel, kind, lily, cruel
> 5. cowardly, dominoes, kangaroo, pansy, tennis

6. during, generous, later, modest, rhinoceros
7. claude, courteous, isaiah, merciful, reasonable
... (Thorndike, 1914, p. 209)

Another set of tests, "Scale for Measuring the Understanding of Sentences and Paragraphs", was designed by Thorndike to measure pupil ability to answer questions about a series of sentences. He stated that:

> Mere word knowledge is much less important than the ability to get the message carried by a continuous passage. Competent judges would rate the latter as from sixty to ninety per cent of the total result to be sought by the elementary school in the teaching of reading. Probably no other one scale for educational measurement is so important as a scale for measuring the understanding of sentences and paragraphs. (Thorndike, 1914, p. 238)

Actually the Scale for Measuring the Understanding of Sentences and Paragraphs was made up of two subgroups. One group of sentences contained narratives or anecdotes. Students were asked to read the sentences and then answer questions:

> In Franklin, attendance upon school is required of every child between the ages of seven and fourteen on every day when school is in session unless the child is so ill as to be unable to go to school, or some person in his house is ill with a contagious disease, or the roads are impassable.
>
> 1. What is the general topic of the paragraph?
>
> 2. On what day would a ten-year-old girl not be expected to attend school?
>
> 3. Between what years is attendance upon school compulsory in Franklin?...(Thorndike, 1914, p. 267)

The other group of sentences contained directions, which were quite simple at the lower levels, but complicated by numerous qualifying conditions at higher levels:

> In these two lines draw a line under every 5 that comes just after a 2, unless the 2 comes just after a 9. If that is the case, draw a line under the next figure after the 5:
>
> 5 3 6 2 5 4 1 7 4 2 5 7 6 5 4 9 2 5 3 8 6 1 2 5
>
> 4 7 3 5 2 3 9 2 5 8 4 7 9 2 5 6 1 2 5 7 4 8 5 6
>
> (Thorndike, 1914, p. 247)

These comprehension questions corresponded to Thorndike's conception of understanding:

> Understanding a paragraph is like solving a problem in mathematics. It consists in selecting the right elements of the situation and putting them together in the right relations and also with the right amount of weight or influence or force for each. The mind is assailed as it were by every word in the paragraph. It must select, repress, soften, emphasize, correlate and organize all under the influence of the right mental set or purpose or demand. (Thorndike, 1917b, p. 329)[1]

After developing the comprehension tests, Thorndike administered them to large numbers of children and conducted careful analyses of errors. From the error analysis, he concluded that mistakes on tests were due to three causes. The first error resulted from mistakes in word definition. Pupils attributed either wrong or inadequate meanings to words in the paragraph or question and developed their answers around this misinterpretation. For example, in the previous paragraph about the rules for school attendance in the city of Franklin (p. 18), some students defined Franklin as a man's name rather than the name of a city. Other students went even further and confused Franklin with a

[1] Thorndike's view of the nature of comprehension was really an outgrowth of his more general connectionistic theory of learning. See Hilgard (1956) Chapter II.

particular man in such answers as "a great inventor." As Thorndike described the errors, "...a word may produce all degrees of erroneous meaning for a given context, from a slight inadequacy to an extreme perversion (Thorndike, 1917b, p. 327)."

Thorndike called the second type of error "over-potency." Over-potency resulted when pupils chose an element such as a fact in the paragraph, a word in the question, or a fact from general experience, attributed undue importance to it, and formulated an answer around it. For example, in the previous paragraph about Franklin, pupils who stated that the topic of the paragraph was "Franklin attends school" gave over-potency to the element "Franklin."

The third type of error--a complement of the second--was called "under-potency." Under-potency referred to mistakenly ignoring the influence of a word in the paragraph or question. Using the example of school days in Franklin again, students were asked, "On what day would a ten-year-old girl not be expected to attend school?" Students demonstrated under-potency of the word "not" in answers such as "when school is in session" or "five days a week (Thorndike, 1917b, p. 328)."

As a result of his investigations, Thorndike made three observations about reading comprehension. First, mental set was very influential in the way students understood selections and answered questions. Second, reading comprehension difficulty could be due to the structure of either the question or selection. Third, a discrepancy could exist between understanding the words and understanding the task. For example, even though a pupil might understand the words in the selection and the question, he might not understand what he is expected to do in order to demonstrate his comprehension. More often than not, the way the comprehension tests were organized made it impossible to establish which of these aspects led the student astray:

> One could in fact make a scale...harder, using just the same paragraph and using questions simply phrased, but demanding the understanding of more and more intricate, subtle, and technical features of the paragraph. Eventually, we may expect to have at least two scales,--one of harder and harder paragraphs or questions or both, each to be read perfectly, the other of a few paragraphs to be read with increasing degrees of fullness and exactitude. The present scale is a mixture. (Thorndike, 1915, p. 460)[1]

As a result of his third observation, Thorndike began to experiment with both verbal (answering a question in narrative form) and action (answering a question by following directions) responses in his paper and pencil tests.[2] Thorndike stated that "measures of ability to understand should be unconfused by ability to express oneself orally or in writing (Thorndike, 1914, p. 227)." He therefore preferred multiple-choice and short-answer questions to the longer, less restrictive essay questions.

Thorndike explicitly stated that his tests were not designed to diagnose skill deficits. A teacher would not know a pupil's specific strengths or weaknesses in reading from a score on these tests. Nor would a teacher know what should or should not be taught in reading comprehension. The tests did not set standards or objectives for

[1] Most current tests still represent a mixture. As yet, there are no valid independent measures of selection and question difficulty; and thus there are no systemized scales that vary the one or the other knowledgably.

[2] See sample items on pp.17,18. The Franklin item represents a narrative response while the drawing lines item represents an action response.

instruction. All a teacher would know from a pupil's score on these tests was how well the pupil could perform on a certain combination of reading test items in relation to many other pupils of corresponding age or grade.

The assumptions underlying the construction of Thorndike's tests were generally that "achievement of paragraph reading is distributed approximately in the form of the so-called normal probability surface..." and "...that the variability of any grade from the fourth to the twelfth... is approximately equal to that of any other (Thorndike, 1916, p. 41)." Thus, items were designated for a specific grade level if they were passed by the major proportion of pupils at that grade level, by a lesser proportion of pupils at the adjacent lower grade, and by a greater proportion of pupils at the adjacent higher grade. However, that did not help teachers establish whether or not pupils could read and understand textbooks or more general types of reading matter. The test scores merely reflected a kind of natural phenomenon:

> What will be achieved as the science of education progresses can not be stated. What should be achieved now if the best known methods were used by the best teachers now available, I will also not try to estimate. What are called "standards" here are simply achievements a little above those actually made in schools under the possibly disturbing conditions of test (sic) by an outsider.
>
> A school whose pupils are able to read as well as this is probably doing better than the general run of schools, but ...it is not achieving enough to enable its pupils to read easily the text-books they are studying, to say nothing of more difficult discussions in newspapers and magazines. (Thorndike, 1915, p. 458)

Thorndike's awareness that only a relative comparison among pupils was possible with his test, did not prevent him from suggesting some objectives for a desirable level of reading achievement; nor did it prevent him from trying to make the test as pure a measure of comprehension as possible. However, despite attempts to isolate the ability to understand paragraphs, he had to reconcile himself to the fact that testing would probably be limited to measuring combinations of factors:

> The scale even when properly administered will occasionally measure a mixture of general stupidity or indolence or mischief with an inability to understand words. Probably no scale could be objective and convenient in use without suffering from this limitation. (Thorndike, 1914, p. 226)

Although it had limitations, standardized group testing as first developed by Thorndike permitted the evaluation of classes, teachers, methods, and schools by more objective criteria than were previously available. For the first time comparisons among school districts, socio-economic and ethnic populations became possible. In part large testing programs were facilitated by Thorndike's introduction of scoring keys and record sheets. He made scoring and tabulating so simple that it could be done very quickly even by non-professionals.

Two more aspects of Thorndike's work should be noted. First, he measured the time factor. He suggested that students should not be given speeded tests since unnecessary anxiety might be produced. However, he thought reading rate was valuable information for the teacher. Generally, he anticipated that older children would work faster than younger ones, and that more intelligent children would work faster than duller ones.

Second, Thorndike identified comprehension with thinking:

> Understanding a spoken or printed paragraph is then a matter of habits, connections, mental bonds, but these have to be selected from so many others, and given relative weights so delicately, and used together in so elaborate an organization that "to read" means "to think," as truly as does "to evaluate" or "to invent" or "to demonstrate" or "to verify." (Thorndike, 1918, p. 114)

To summarize, Thorndike made significant contributions to the validity, construction, design, administration and scoring of reading comprehension tests. Thorndike

1. strived to specify the kind of comprehension being tested
2. introduced standardization and norming procedures
3. identified discrepancies among selection-question-response difficulty
4. demonstrated appropriate use of the time factor
5. facilitated simultaneous testing of many students
6. developed quick and economical scoring procedures.

Innovations in Testing

Since E. L. Thorndike the most dramatic changes in the development of reading comprehension tests have been technological. Currently published standardized tests are normed on as many as 260 school systems in 50 states and on 850,000 pupils (Kelly, et al, 1964a).

Widely used tests include about 15,000 questions in their experimental forms (Kelley, et al, 1964a). Percentages are computed for the number of children choosing each multiple-choice distractor. On the basis of these percentages, "item profiles" are constructed:

> These item profiles were considered one of the most important indices of item validity, and considerable weight was attached to them in the selection of items for the final forms. Results of this item tryout permitted identification of ambiguous items, of items either too easy or too difficult for the grades for which they were intended, and items unsatisfactory in other respects. Such items were eliminated from consideration for retention in the final forms.... (Kelley, et al, 1964a, p. 26)

In addition to improvements in norming procedures, considerable refinements have been made in the mechanics of test administration and scoring. Time restrictions are investigated during the norming procedure, and the limit selected represents the amount of time required by a specified percent of the norming population to complete the prescribed task. Statistical innovations, particularly the development of stanines, permit score comparison by equal units. Computers make possible the scoring of answer sheets rapidly and at minimal cost (Harcourt, Brace and World, 1968; California Test Bureau 1968).

Another major innovation has been the introduction of cloze procedure.[1] Cloze procedure is a systematic deletion of every nth word in a passage of prose. Usually every 5th word is deleted. Pupils are asked to fill the blanks. Mostly, only an exact replacement of the deleted word is marked correct (Taylor, 1953; Tremont, 1967;

[1] A more detailed analysis of cloze as a test of comprehension is given by Tremont (1967, p. 50-66) and Bormuth (1969b).

Bormuth, 1969b).[1] First introduced by Taylor in 1953:

> Cloze procedure derives its name from the "closure" concept of Gestalt psychology. Just as there is an apparent human tendency to "see" a not-quite-complete circle as a whole circle - by "mentally closing the gap" and making the image conform to a familiar shape - so does it seem that humans try to complete a mutilated sentence by filling in those words that make the finished pattern of language symbols fit the apparent meaning. (Taylor, 1957, p. 19)

Cloze procedure has solved some problems in testing while at the same time creating new ones. The cloze test is considerably simpler to construct than a question test. It also eliminates the interference of question content and structure in tests of comprehension (Bormuth, 1966, 1967; Simons, 1970). Unfortunately, however, it is not always clear what cloze is measuring.[2] Taylor (1957) found cloze scores to be

[1] This type of scoring penalizes pupils who may put down a perfectly valid, although not identical, answer. Thus, even though these pupils in fact demonstrate comprehension, their performance is rated as inadequate. Cloze shares this problem wirh all other comprehension tests that require one answer where many may be equally appropriate. Thus, accuracy in testing is sacrificed for simplicity in scoring.

Trenaman (1967) discusses in somewhat greater length how people may validly differ in their understanding of a language communication.

[2] For example, Tremont (1967, p. 66) suggested cloze "may be an excellent test for measuring the interrelationships among ideas;" and thus better than most traditional tests, which he concluded, "measure word meanings, literal meanings of sentences, and only occasionally consider measuring relationships among ideas." Bormuth (1969b, p. 365) concluded that cloze tests "measure skills closely related or identical to those measured by conventional multiple-choice reading comprehension tests." And, Simons (1970, p. 14) concluded that cloze is a "better measure of comprehension because...it appears to be measuring fewer extraneous aspects of cognitive functioning than traditional tests do."

Generally, deciding what best measures comprehension seems very much dependent on how reading comprehension is defined. The present lack of information with respect to cognitive functioning in the case of reading makes it very difficult to establish which cognitive functions are or are not extraneous to reading.

a dependable index of mental ability, of previous knowledge, and of information known after reading a given prose passage.[1] Furthermore, he established that the parts of speech deleted determined the difficulty of the tests. Deletions of every nth noun, verb and adverb created the test of greatest difficulty. Deletions of every nth adjective or preposition proved to be of intermediate difficulty. Deletions of every nth auxiliary verb, conjunction, pronoun or article created the simplest test. Taylor also constructed what he called "any" tests, by deleting every nth word irrespective of its part of speech. The "any" type of cloze was easier to construct than the other forms. It also proved most satisfactory in providing stable results and discriminating among testees. Furthermore, Taylor (1957) established that his most difficult test (deletion of every nth noun, verb or adverb) was the best indicator of previous knowledge.

Current widely-used comprehension subtests are not designed by regularly deleting every nth word. Rather 1, 2 or 3 words in a sentence or paragraph are deleted with no rationale or explanation given by the test-authors. An informal review by the present writer of cloze-like blanks in selected tests of reading comprehension revealed that deleted words were generally nouns, verbs, or adverbs. Thus, if Taylor's (1957) findings may be applied here, these tests would be testing previous knowledge.

[1] Taylor (1957) used a technical report on Air Force supply systems as his reading selection. The subjects of his study were 152 Air Force trainees. However, his findings about the difficulty of tests constructed by deleting the various parts of speech seem to have more general significance. It seems to the present writer that the relative difficulty of these types of tests may remain constant both for different types of prose passages and for different groups of readers.

Another innovation in testing has been the almost universal acceptance of multiple-choice. Rather than writing their own answer, pupils are asked to respond with one out of four or five answer choices. Generally one choice is the correct answer and the other choices act as distractors. The benefits of this innovation in testing are quick and objective scoring. A disadvantage of this innovation is that the pupil's answer to a multiple-choice item "may be influenced by the distractors from which he has to choose just as much as by the question part of the item (Schlesinger and Weiser, 1970, p. 569)." Bormuth (1966, p. 82) also contended that "it is notoriously easy to vary the difficulties of these tests simply by changing the alternatives to the question." Since little is known about distractor combinations, little is known about what a pupil must do to choose the correct answer. Guttman and Schlesinger (1967) have begun studying the types of errors pupils make in choosing distractors. They have concluded that there are consistencies in types of errors pupils make, and that identification of these consistencies may prove diagnostically useful.

After studying current developments in educational testing, P. E. Vernon concluded that:

> Whatever the subject matter - English, social studies or natural sciences - they tend to take the form of complex reading comprehension tests, and they therefore appear to depend partly on the students' facility in understanding the instructions and coping with multiple-choice items. (Vernon, 1962, p. 269)

Vernon supported his conclusion by pointing out that the correlations between tests aimed at different mental functions or different school subjects were extremely high.[1] He further hypothesized that the differences among tests may "be due merely to the imperfect reliability of the contrasted tests (Vernon, 1962, p. 270)."

Investigations into the Nature of Reading Comprehension

The review of the literature thus far has focused on the development of reading comprehension tests. However, this development was not isolated. Rather, it was related to the studies of scholars from many disciplines who investigated reading comprehension for different purposes and by different approaches.

The three most prevalent methodologies used to study comprehension seemed to evolve in a historical sequence. The first type of study was philosophical investigation based on intuition and logical analysis. Topics of concern included the goals of reading comprehension, means to attain these goals, relationship of thought to language, and relationship of reading matter to understanding. The treatise of Locke (1697) on the Conduct of the Understanding, the writing of Stewart (1811) in Philosophical Essays, and the analysis of Smart (1855) in Thought and Language are examples of early philosophical works concerned with reading comprehension.

[1] Vernon quoted the mean intercorrelation for five of the Iowa Tests of Educational Development (basic social concepts, reading in social studies, reading in natural science, interpreting literary materials and vocabulary) among 9th and 12th grade students of .716 while the tests had a mean reliability of .905 (Vernon, 1962, p. 270).

Experimental psychology provided a second approach to the study of reading comprehension. Philosophical investigations, such as the work of Richards (1929), and Wittgenstein (1958) continued, but the newer experimental techniques seemed more prevalent.

The experimental studies were characterized by the testing of given hypotheses about reading comprehension. Criteria had to be specified that would either substantiate or refute the hypotheses. Thus, an increased interest developed in specifying the desired goals of comprehension, as well as criteria for determining the occurrence of comprehension. Ingenious machines to record the rhythm and sequence of visual movements were developed. Readers' introspections were noted and analyzed. Investigations of empirical phenomena--observing behaviors of large groups of readers--were also conducted. The works of Huey (1908, reissued 1968), Thorndike (1914-18), James (1928) and Skinner (1937) exemplify the experimental period. Generally, the experimental approach resulted in testing--substantiating or refuting--some of the speculations proposed earlier by philosophers. The experimental approach also created the need for better objective evaluation of reading data. Statistics thus provided the third approach to studying reading comprehension. As before, previous approaches were not rejected. Philosophical and experimental studies continued. Occasionally studies of comprehension incorporated all three approaches.[1] However, statistical studies became most prevalent.

[1] Levin and Williams (1970) present an interesting combination of recent studies of reading. Although the studies incorporated in their book are largely of an experimental nature, both philosophical and statistical approaches are represented.

The statistical analyses were generally of two types. One type was process-oriented. Studies of this type analyzed and interpreted factors and correlations that reflected the comprehension process itself (Gans, 1940; Langsam, 1941; Davis, 1944, Hall and Robinson, 1945; Thurston, 1946; Anderson, 1949; Johnson, 1949; Sochor, 1959; Alshan, 1964; Holmes and Singer, 1966; Davis, 1968; Trenaman, 1967). The studies were not always based on widely-used standardized tests of reading comprehension. For example, Davis in 1944 developed his own questions by making "...a careful survey...of the literature to identify the comprehension skills deemed most important (Davis, 1968, p. 504)."

Conclusions drawn from process-oriented studies have produced inconsistent results. Some investigators concluded that reading comprehension had numerous factors while others concluded that comprehension was one general factor.[1] As might be expected, the resultant factor or factors reflected the structure and content of the tests as well as the statistical treatment.[2] Sometimes, the resultant factors were almost identical to the criteria used for devising test items. Often, in factor analytic studies, the criteria as well as the outcome factors suffered from confusion of requirements for reading, prerequisites for reading, procedures for teaching reading and skills or

[1] For a lengthier discussion of the controversy among the factor analytic studies in reading comprehension, see Hunt (1957).

[2] Davis (1944, p. 185) stated, "Unless these tests provide reasonably valid measures of the most important mental skills that have to be performed during the process of reading, the application of the most rigorous statistical procedures can not yield meaningful and significant results. The importance of this point can hardly be overstated."

abilities used in reading (Strang, 1965; Robinson, 1966). Despite the disagreements among studies a few consistent findings did appear. Most of the studies that identified a number of comprehension factors seemed to agree on four: vocabulary factor, interrelationship among ideas (represented by words in context) factor, abstract reasoning factor, and specific content field factors (Jenkinson, 1968; Simons, 1970).

The second type of statistical analysis investigated the relationship of factors outside the comprehension process itself to comprehension as measured by standardized achievement tests. Among the variables studied were age, sex, race, socio-economic status, personality traits and intelligence (Bleismer, 1954; Gates, 1961; Vehar, 1962; Cooper, 1964; Chandler, 1966; Coleman, 1966; Harootumian, 1966; Neville, Pfost and Dobbs, 1967; Dykstra, 1968). Because of the differences in the reading tests and in the size and composition of the samples used, few valid generalizations could be drawn from all these correlational studies. Two generalizations that seemed consistent, however, were the positive correlations of tested reading comprehension with tested general intelligence and with socio-economic factors. Both of these correlations increased with increasing age.

Roger Farr (1969) in his comprehensive study, Reading: what can be measured? reviewed and synthesized major contemporary research in reading comprehension. He discussed the continued controversies in measuring comprehension: emphasizing reading rate vs. comprehension power, permitting reference to the reading selection vs. removing the selection, controlling for previous knowledge vs. ignoring it, establishing purposes for reading vs. not doing so, testing solely for syntax vs. testing for many "skills," varying lengths of reading selections vs. keeping constant lengths, and controlling for personality traits vs. ignoring them. Farr (1969, p. 56) concluded that "there is still a lack of understanding about the basic aspects of reading comprehension."

The review of the philosophical, experimental and statistical studies of reading comprehension led the present writer to conclude that there is considerable interdependence among them. The results of "objective" experimental and statistical analyses are generally colored by intuitive and subjective criteria. In most experimental and statistical studies, tests were used as the criteria of reading comprehension. Validity of the tests was judged by the researcher and test-author.[1] Often the researchers' and test-authors' conception of comprehension resulted from one or a combination of earlier philosophical positions. Tests and experimental designs tended to reflect at least in part, one or another philosophical orientation. Subsequently, differences among conclusions resulting from statistical analyses and "experiments" reflected to some extent the corresponding differences in philosophies, and thus were not totally "objective."

Understandably, Farr suggested that "The most pressing research need in measuring comprehension is for a clear understanding of the

[1] Guilford (1946, p. 437) points out that "Even sophisticated judgment often goes astray on decisions as to what a test measures. A test designed to measure common sense judgment when factor analyzed turns out to be a test of mechanical experience. A test designed as a reasoning test is found to be one of numerical facility, when analyzed.... The moral of it is that in test construction..., things are not always what they seem."

nature of reading comprehension (Farr, 1969, p. 64)." Finding where to begin such an undertaking is no easy task. Miller (1967, p. 90) pointed out that "No psychological process is more important or difficult to understand than understanding, and nowhere has scientific psychology proved more disappointing to those who have turned to it for help." Therefore, in the present writer's opinion, it may be most practical to start with an analysis of empirically constructed (standardized and normed) comprehension tests. These tests are the accepted criteria of reading comprehension. There is, however, no clear understanding yet of what these tests are measuring. An understanding of widely-used reading comprehension tests may lead to a better understanding of what is currently being called reading comprehension.

CHAPTER III

Analysis of Standardized Reading Comprehension Tests

Introduction

Standardized reading comprehension tests have been, for many years, the accepted method for evaluating reading comprehension. Because of this these tests offer a source of empirical data which, if analyzed properly, may improve the understanding of reading comprehension.

The elaborate norming procedures and item analyses that characterize standardized tests establish an empirical scale of comprehension difficulty.[1] Test-authors select items for a given grade level only if the items successfully differentiate between the high and low achievers at that level and also between that level and adjacent levels. To date, no one seems to know why or how these items differentiate. However, the fact that the test items do empirically discriminate among pupils and grades implies strongly that comprehension items reflect an underlying structure or sequence of reading comprehension tasks. A systematic analysis of reading comprehension test items therefore, may reveal this structure.

[1] See Chapter II, p. 17 of this dissertation for a more detailed account of the procedure used in developing tests.

Objectives

The objectives of this systematic analysis of reading comprehension tests were to:

1. characterize the nature of reading comprehension as tested at three grade levels (grades 1-2, 4-6, and 9-14); e.g., Are there differences in what pupils have to do or know in order to demonstrate reading comprehension on tests at the different grade levels? What kinds of changes occur from one grade level to the next?
2. characterize the nature of reading comprehension as tested by different test batteries; e.g., Are there differences in what pupils have to do or know in order to demonstrate a given level of reading comprehension in different test batteries? What kinds of differences exist?
3. identify factors that may contribute to difficulty in tested comprehension; e.g., What are the factors that make one test question more difficult than another; or one test more difficult than another?
4. characterize the nature of tested reading comprehension; e.g., What do comprehension tests test? What does a pupil have to <u>do</u> in order to demonstrate reading comprehension on these tests? What does a pupil have to <u>know</u> in order to demonstrate reading comprehension on these tests?

Tests

Reading comprehension subtests selected for this study were the California Achievement Test (1963), form W, comprehension/ interpretation subtest (CAT), Gates-McGinitie Reading Test (1965, 1969), form 1, comprehension subtest (GMRT), and Stanford Achievement Test (1964, 1965), form X, paragraph meaning/reading subtest (SAT).[1]

The subtests chosen were designed by their authors to measure understanding and comprehension (see Table 2, p.13). The test batteries (CAT, GMRT, SAT) were selected on the basis of five criteria:

[1] The names of the comprehension subtests varied among the batteries. The CAT subtest at the lower primary level was called comprehension. However, at the higher levels the comprehension subtest on the CAT was broken down into three parts: interpretation, following directions and reference skills. The "interpretation" part was analyzed in this study since it most closely resembled the comprehension subtests of the other batteries. The "following directions" part of the CAT comprehension subtest was made up of short passages giving math, history or science information, a direction requiring the identification or application of the given information and four answer choices. The "reference skills" part included questions usually requiring knowledge of reference materials (e.g. dictionaries, maps, graphs) and four answer choices.

The SAT comprehension subtest was called paragraph meaning at the lower levels, however, on the highest level (high school) the subtest name was "reading." In both the SAT and GMRT the highest level tests were published later than the rest of the battery. Thus, while most of the SAT battery was published in 1964, the High School test was not published until 1965. Similarly with the GMRT, most of the battery was published in 1965, but the highest level test, Survey F, was not published until 1969.

1. that the reading comprehension subtests be comparable;[1]
2. that the tests be standardized;
3. that the tests be normed at relatively corresponding grade levels;
4. that the tests be widely used in the United States;
5. that the tests be distributed by different publishers.

Three grade levels were chosen within each test battery in order to observe the progression of reading comprehension difficulty. The lowest level tests were for grades 1 and 2, the intermediate level for grades 4 through 6, and the highest level tests were for grades 9 through 14.[2]

[1] Although the subtests were generally comparable in format, a number of differences existed. These differences were most evident at the lowest grade level. The lowest level CAT and GMRT contained a number of "direction" items. In these items answer choices followed the "direction", but there was no reading selection as with most items on the lowest level SAT subtest and on all higher level subtests. In addition, the first grade level GMRT presented picture answer choices while all other subtests analyzed had word answer choices. Also, the first grade level CAT included one open-ended (no choices) item and one "direction" that consisted of copying the initial letter of a word; and, two were mutilated words that had to be fixed.

[2] The grade levels for which the subtests from different batteries were intended by their authors were not entirely consistent. Specifically, at the lowest grade level, the GMRT (Gates and MacGinitie, 1965a, p. 1) was intended for first grade only, while the CAT (Tiegs and Clark, 1963c, p. 1) and the SAT (Kelley, et al, 1964c, p. 1) were intended for grades 1 and 2. At the intermediate level the SAT (Kelley, et al, 1964b, p.2) authors wrote two tests: one for grade 4 and one for grades 5 and 6. The CAT (Tiegs and Clark, 1963b, p. 1) and the GMRT (Gates and MacGinitie, 1965b, p. 1) authors published only one test for grades 4 through 6. At the highest level, the CAT (Tiegs and Clark, 1963a, p. 1) was intended by its authors for grades 9 through 14, while the GMRT (Gates and MacGinitie, 1970, p. 1) was intended for grades 10 to 12, and the SAT (Gardner, et al, 1965, p. 1) was intended for grades 9 to 12.

The tests analyzed generally had similar formats. Most comprehension items consisted of a reading selection, a question, and choices. The reading selection usually consisted of a sentence, a paragraph, or a number of paragraphs. The selection either contained a number of cloze-like blanks, or was followed by one or more separate questions. Four or five answer choices were generally provided by the test-author. Pupils were required to choose the "choice" which correctly filled the "blank" or answered the question.

Table 3 summarizes the number of selections, questions, and choices analyzed in each subtest and at each level.[1] A total of 165 selections, 455 questions and 1902 choices were analyzed in all.

Analyses

Only a short introduction to the types of analyses conducted in this study will be presented here. In Chapter IV and V more specific descriptions will be given of the analysis procedures.

Reading selections, questions, and choices were analyzed in two ways. First a Dale-Chall (1948) and Spache (1953) readability analysis was made of each selection, question and choice. These and similar readability formulae are used to appraise objectively the relative difficulty of basal readers, textbooks, encyclopedias, newspapers, and standardized tests (Chall, 1956, p. 89). The predictions of reading difficulty of the Dale-Chall (1948) and the Spache (1953) readability formulae are based on counts of the number of difficult words in a reading selection and also the average number of words in the

[1] Where the reading selection contained cloze-like blanks, the sentence containing the blank was counted and later analyzed as the question.

Table 3

The Number of Selections, Questions and Choices Analyzed in this Study by Test and Level

Test	Level Lowest (Grades 1-2)	Intermediate (Grades 4-6)	Advanced (Grades 9-14)	All Levels
CALIFORNIA ACHIEVEMENT TEST, Form W				
Selections	4	3	5	12
Questions	15	30	45	90
Choices	38	120	180	338
GATES-MACGINITIE READING TEST, Form 1				
Selections	16	21	21	58
Questions	34[a]	52	52	138
Choices	136[b]	260	260	656
STANFORD ACHIEVEMENT TEST, Form X				
Selections	33	24 25[c]	13	95
Questions	38	60 64	65	227
Choices	152	240 256	260	908
ALL TESTS				
Selections	53	73	39	165
Questions	87	206	162	455
Choices	326	876	700	1902

[a] Unlike other questions, some questions on this subtest did not refer to a reading selection.

[b] These were picture choices rather than word choices as in other tests.

[c] The Stanford Achievement Tests had one test for grade 4 (Intermediate I) and another for grades 5 and 6 (Intermediate II). Both these tests were analyzed since the other batteries at that level were for grades 4 through 6.

sentences of the selection (Chall, 1958b; Klare, 1965).[1]

The second type of analysis was designed especially for this study. Three judges rated each selection, question and choice. Selections were rated according to topics or general subject areas (e.g. history, science, etc.). Questions were rated according to the relationship of the correct answer choice to the word(s) presenting the information in the selection (e.g. same word in a different context, grammatically different, etc.). Wrong choices, called distractors, were rated according to their relationship to the selection (e.g., in the selection or not), to the question (e.g., grammatical answer or not), and to the correct choice (e.g., coordinate, superordinate, etc.)

The data obtained from the two analyses described above were studied for clear and/or consistent trends rather than for specific statistically significant differences. Since there were often wide discrepancies in the number of cases to be compared and since a great number of statistical relationships would have been explored, the assumptions underlying most statistical tests of significance were or would have been violated.[2]

[1] Difficult words in these formulae are identified by their absence on given lists of easy words. The Dale-Chall formula (1948) uses the Dale List of 3000 Familiar Words and the Spache (1953) formula uses Clarence H. Stone's Revision of the Dale List of 769 Easy Words.

[2] The problem of carrying out many non-independent tests of significance is discussed by Kendall and Stuart (1966, v. 3, p. 40). For example, the probability of finding a significant difference where none exists is approximately equal to the product of the level of significance (α) and the number of tests of significance computed (K) or Probability $= \alpha K$. This formula is for independent sets of data. However, it provides an approximation for interrelated data such as is present in this study. Consequently, even if a significance level of .01 is used in each test of significance, the probability of finding a significant difference where none exists in 25 tests of significance would be approximately .01 x 25 = .25, or 1 in 4, rather than the anticipated 1 in 100.

CHAPTER IV

Readability Measurement

Introduction

As Lorge (1949, p. 86) defined it: "The concept of readability involves the idea of understanding printed material." Attempts at measuring readability have been traced back hundreds of years (Klare, 1963).

More recently, Chall (1958b, p. 156-158) formulated seven major generalizations about readability measures from the "fundamental methodological research in readability":

1. a variety of factors contribute to reading difficulty... content, stylistic elements, format, and organization....

2. ...only stylistic elements have been amenable to reliable quantitative measurement and verification.

3. of the diverse stylistic elements...only four types can be distinguished: vocabulary load, sentence structure, idea density, and human interest.

4. of the four types of stylistic elements, vocabulary load (diversity and difficulty) is most significantly related to all criteria of difficulty so far used. Vocabulary difficulty has to do with the reader's understanding of the individual words.... Vocabulary difficulty has been measured either by reference to a word list or by word length....

5. almost every study found a significant relationship between sentence structure and comprehension difficulty. The most popular method of estimating sentence structure is by sentence length....

6. readability formulas measure idea density only indirectly through the percentage of prepositional phrases and, less often, through the percentage of different content words.... Prepositional phrases have less potent influence on difficulty than either vocabulary difficulty or sentence structure. They add little to the over-all predication of difficulty, once some measure of these two factors is included in a formula.

7. human interest has been measured by number of personal pronouns, persons' names, and nouns denoting gender.... However, these measures add little to a readability formula, once vocabulary difficulty and sentence structure are used.

Chall (1958b)and Klare (1963) described the many readability formulae that had been developed through different combinations and weightings of the stylistic elements described above. Klare (1963) identified 31 formulae by 1960. Currently, new advances in linguistic theory, particularly the work of Noam Chomsky (1965), have prompted new developments in measuring readability.[1]

The present study explored the concentration of the most potent predictors of relative "language" difficulty (described above) in selections, questions and choices. Toward this end, the Dale-Chall (1948) and the Spache (1953) formulae were used. Two formulae were necessary since each formula had grade level limitations as do most formulae. The Spache formula was designed to rank reading matter from the grade 1 through the grade 3 level. The Dale-Chall formula was designed to rank reading matter from the grade 4 through the college graduate level. Thus the Spache is a more appropriate measure for the lower level tests; while the Dale-Chall is a more appropriate measure

[1] Bormuth (1967) discusses the areas of advancement in readability research: the use of cloze, developments in linguistic theory, and the use of finer statistics.

for the higher level tests.

Both formulae consist of a measure of vocabulary load (number of difficult words not on a given list) and a measure of sentence structure (average sentence length). Therefore, they seem more comparable to each other than to readability formulae consisting of other measures such as word length, or a different kind of readability element such as human interest. However, some differences do exist between the two formulae, which may account for their appropriateness to different grade levels. The Dale-Chall formula is based on the Dale List of 3000 Familiar Words; while the Spache formula is based on Clarence H. Stone's Revision of the Dale List of 769 Easy Words. Furthermore, while the Dale-Chall formula weighs the number of times a difficult word occurs, the Spache formula counts difficult words only once.

Another reason for using the Spache and the Dale-Chall formulae in the present study was that they appeared to be frequently used in appraising educational materials. A demonstration of their popularity was the frequent use of these formulae in appraising the difficulty of randomly selected comprehension skill builders reviewed in Chapter I (see Table 1, p. 5). Finally, the Dale-Chall formula seemed to be among both the formulae that correlated most highly with other readability formulae, and the formulae that gave the most valid grade scores for juvenile fiction of intermediate difficulty (Chall, 1958b, p. 164).

The readability analysis also provided a means of comparing relative difficulty of test items, as rated by readability formulae, and the empirical difficulty scores of the items provided by the test publishers. Empirical difficulty refers to the percentage of pupils answering a given test item correctly. Test items generally included a reading selection, a question about that selection and answer choices to the question.

An exploration of the distribution of selection, question and choice readability scores follows. Comparisons were made among reading comprehension tests at three grade levels (1-2, 4-6, and 9-14) and in three test batteries (California Achievement Test, 1963; Gates-MacGinitie Reading Test, 1965, 1969 and Stanford Achievement Tests 1964, 1965). Also reviewed were the relationships of these readability scores to each other and to the empirical item difficulty scores.

Procedure

Readability Scores

Two Harvard doctoral students independently counted the following variables for each:

1. reading selection
   a. the number of words in the selection
   b. the number of sentences in the selection
   c. the number of words not on Clarence H. Stone's Revision of the Dale List of 769 Easy Words (non-Spache)
   d. the number of words not on the Dale List of 3000 Familiar Words (non-Dale-Chall)
2. question
   a. the number of words in the question

b. the number of words not on Clarence H. Stones Revision of the Dale List of 769 Easy Words (non-Spache)

c. the number of words not on the Dale List of 3000 Familiar Words (non-Dale-Chall)

3. choice

a. the number of words in the choice

b. the number of words not on Clarence H. Stone's Revision of the Dale List of 769 Easy Words (non-Spache)

c. the number of words not on the Dale List of 3000 Familiar Words (non-Dale-Chall)

When both investigators finished all the items in a given test, they compared their counts. If counts for any selection, question or choice conflicted, both investigators again counted that part of the item independently. Results were compared again. This procedure was repeated until agreement was reached.

The "word counts" provided the data which were punched onto IBM cards. Further computation was conducted on the IBM 360/65.[1] Randomly selected computations were checked with the Olivetti Programma. The following scores were computed for each:

1. Reading selection

a. average sentence length in the selection

b. Spache ratio - number of non-Spache words in the selection divided by the total number of words in that selection

c. Dale-Chall ratio - number of non-Dale-Chall words in the selection divided by the total number of words in that selection.

[1] All the scores below were computed according to the specifications of the Dale-Chall (1948) or the Spache (1953) formula.

d. Dale-Chall raw score - Dale-Chall ratio (c above) multiplied by the constant 0.1579 and added to average sentence length multiplied by the constant 0.0496. A constant, 3.6365, was added to this sum.[1]

e. Spache grade score - average sentence length multiplied by the constant 0.141 and added to the Spache ratio (b above) multiplied by the constant 0.086. A constant, 0.839, was added to this sum.

f. Dale-Chall grade score - Dale-Chall raw score (d above) was converted into corrected grade levels from a table.[2]

2. question

a. Spache ratio - number of non-Spache words in the question divided by the total number of words in that question

b. Dale-Chall ratio - number of non-Dale-Chall words in the question divided by the total number of words in that question

[1] The constants for both the Dale-Chall raw score and the Spache grade equivalent resulted from a multiple-regression technique. For further information about the regression technique see Chall (1958b) and Klare (1963).

[2] The table of corrected grade levels provided by Dale and Chall (1948) consists of ranges. For example, a raw score of 5.0 to 5.9 corresponds to a grade level range from 5th to 6th grade. For purposes of simplifying the computation, the midpoint of this range was used in computing means and standard deviations for this study. Thus, for the above range a grade equivalent of 5.5 was assigned to a raw score between 5.0 and 5.9.

3. choice

   a. Spache ratio - number of non-Spache words in the choice divided by the total number of words in that choice

   b. Dale-Chall ratio - number of non-Dale-Chall words in the choice divided by the total number of words in that choice

The grade scores and average sentence lengths were not computed for the questions and choices because they seemed inappropriate for two reasons. First, the test questions in this study were usually no longer than one sentence and the choices were often only one word, providing essentially no data from which to compute average sentence length. Second, the readability formulae generally were standardized on reading selections of approximately 100 words in length (Chall, 1958b, p. 171). Consequently, conclusions based on average sentence length where none existed, and on grade scores computed with multiple regression coefficients and constants obtained from 100 word samples would have been either extremely tentative or possibly even meaningless.[1] Furthermore, although formulae have been generally accepted

[1] Coefficients and constants of readability formulae result from the particular data sample used in the multiple regression analysis. Another data sample would produce different coefficients (Kendall and Stuart, 1966, v. 2, p. 355). However, formulae may be used for similar samples. There seem to be two types of similarity. One is the content of the reading selection. The content of materials appraised by readability formulae should be similar to the content of the materials on which the formulae were standardized. Both the Dale-Chall and the Spache formulae were standardized, in part, on general type of school reading matter (Chall, 1958b, p. 39). The second type of similarity is the length of the selection. The length of materials appraised by readability formulae should correspond to the length of the materials on which the formulae were standardized - usually 100 words. Chall (1958b, p. 171) concluded that relative difficulty and especially grade placement determined for much shorter reading selections "should be considered tentative."

as valid estimates of relative difficulty of reading matter, their determination of exact grade-level difficulties have been questioned for a long time (Chall, 1956, p. 99). However, the more basic readability factors such as number of words, or number of non-Spache and non-Dale-Chall words, etc. may appraise the relative difficulty of questions and choices adequately.

Difficulty Scores

Difficulty scores-- percent right answers to each test question-- of the standardization population were requested by mail from the publishers of the tests analyzed.[1] Scores of the standardization population were requested because it seemed that these difficulty scores were a by-product of the standardization and norming procedure. Consequently the difficulty scores for the standardization population might have been most readily available. In addition, since most test authors presented descriptions of the standardization population in published technical reports, the need for gathering such descriptive information could have been eliminated. Furthermore, standardization populations usually consisted of carefully stratified national samples that were expected to represent the national school population reasonably well (California Test Bureau, 1957, p. 12;

[1] Standardization and norming were conducted simultaneously. Possibly therefore the authors seemed to use the terms interchangeably (California Test Bureau, 1957; Gates and MacGinitie, 1965d; Kelley, et al, 1966).

Kelley, et al, 1966, p. 9).[1]

After many months, much correspondence and many telephone conversations, it appeared that the difficulty scores for standardization populations were not readily available. Only the SAT published difficulty scores of their standardization population for the 1-2 and 4-6 grade level testsin the Technical Supplement (Kelley, et al, 1966, p. 46-53). The SAT published difficulty scores for a national pre-standardization try-out population at the high school level (Gardner, et al, 1965, p. 16). The GMRT difficulty scores for all levels were from a "nationwide tryout that involved more than 25,000 pupils (Gates and MacGinitie, 1965d, p. 2)." The CAT difficulty scores for all levels were from a nationwide pre-revision investigation.[2]

[1] The California Test Bureau (1957, p. 12) controlled for geographic region and community size in selecting standardization samples. With these controls the performance of the samples of pupils drawn was expected to be an "accurate estimate of the performance of the total pupil enrollment in elementary and secondary schools in the United States."

In norming the Stanford Achievement Test

> The distribution according to region and size of system was established in such a way that the norm sample would duplicate these characteristics for pupils in average daily attendance in public and private schools throughout the country...Public schools (integrated, segregated white and segregated Negro), private non-sectarian and private sectarian schools were included in the sample. (Kelley,et al, 1966, p. 9)

The authors of the Gates-MacGinitie Reading Test also reported careful selection of the standardization population "on the basis of size, geographic location, average educational level, and average family income (Gates and MacGinitie, 1965d, p. 2)."

[2] The California Test Bureau was planning a revision of the CAT in 1968. In order to establish which items should be retained from the old form of the tests, a review of item difficulties was undertaken. The sample of item difficulties was taken from tests sent to the California Test Bureau for scoring by school systems around the country that were using the CAT (telephone conversation with Dr. W. E. Kline, Managing Editor, California Test Bureau, 1971).

Table 4 presents the nature of sample and the grades for which difficulty scores were available, the size of the sample, and the year of testing.[1]

Although samples used to provide the difficulty scores were relatively large (see Table 4) and were also usually stratified according to geographic region and community size (Gardner, et al, 1965, p. 16; MacGinitie correspondence, 1970; Kline conversation, 1971), there were no claims by authors that these samples represented national performance as well as the larger, more carefully selected standardization samples.

Treatment of Data

Means, standard deviations, minimums, maximums and ranges of each readability score (see pp. 45-48) were computed for all selections, questions and choices of the 10 tests at the 3 levels and in the 3 batteries analyzed. Means, standard deviations, minimums, maximums and ranges of difficulty scores were also computed for the 10 tests. These computations summarized the distribution of readability and difficulty scores in individual tests, at the three grade levels and in the three test batteries.

In order to explore the correlations that existed among selection, question and choice readability scores and between readability and

[1] Test publishers did not usually have item difficulty data available for each grade for which the test was intended. In some tests all grades were represented at one level (e.g. CAT Elementary - grades 4,5 and 6), but not at another (e.g. CAT Advanced - grades 9, 10 and 12 only). The difficulty scores for most tests were given by grade. GMRT, Survey F difficulty scores were for a combination of grades--10, 11 and 12.

Table 4

The Nature and Size of the Sample, Date of Testing and Grades for which Difficulty Scores were Obtained

	California Achievement Test	Gates-MacGinitie Reading Test	Stanford Achievement Test
	Interpretation Subtest[a]	Comprehension Subtest[b]	Paragraph Meaning Subtest[c]
Nature of Sample	pre-revision investigation	pre-standardization try-out	standardization population pre-standardization try-out[d]
Size of Sample[e]			
Lowest	137	800	1000
Intermediate	399	800	1000
Advanced	387	300	600
Date Tested	1968	1964, 1968[d]	1963
Grades			
Lowest	1,2	1	1
Intermediate	4,5,6	4,6	4,5,6
Advanced	9,10,12	10-12	9,10,11,12

[a] Information obtained in telephone conversation with Dr. W. E. Kline, Managing Editor, California Test Bureau, 1970.

[b] Information obtained in telephone conversation and by correspondence with Professor W. MacGinitie, Co-author, GMRT, 1970.

[c] Kelley, et al (1966, p. 9, 46) and Gardner, et al (1965, p. 9, 16).

[d] Item difficulty scores for advanced level tests were sometimes given for a different size sample and for a different test date than other difficulty scores in the battery.

[e] Size of Sample refers to the number of pupils tested per grade.

difficulty scores, a common unit of measure was established.[1] The number chosen as the common unit corresponded to both difficulty scores and questions. For example, in establishing the common unit with the selections, all the data on a given selection (see pp. 45-47) were duplicated for each question referring to that selection. Table 5 presents the mean, minimum and maximum number of questions that accompanied a given reading selection in each comprehension subtest analyzed. The number of questions accompanying a given reading selection differed among grade levels and among tests. For example, reading selections on the CAT Lower Primary were accompanied by a mean number of 2.5 questions; while those on the CAT Advanced had an average of 9 questions each. The GMRT Level A, on the other hand, had only one question per selection; thus no weighting was needed.

In order to establish a common unit with choices, the converse procedure was undertaken. A given score (e.g. number of words in the choice, see pp. 46 and 48) was added for the four or five choices that accompanied a question. The result of this duplicating and merging of data was a common unit for selection, question, choice and difficulty scores with which Pearson product-moment correlations were computed.

The common unit was also used to compute selected weighted means (Tables 13-18, and 20 in Appendix A). The weighted means

[1] Table 3, p.40 , presented the number of selections, questions and choices analyzed. Usually there were a number of questions that referred to one selection, and four or five choices that referred to one question.

Table 5

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Questions Accompanying a Reading Selection by Test

Test	Number of Selections	Mean number of Questions	Standard Deviation	Minimum number of Questions	Maximum number of Questions	Range
CAT						
1. Lower Primary	4	2.50	1.73	1.00	4.00	3.00
2. Elementary	3	7.00	1.73	5.00	8.00	3.00
3. Advanced	5	9.00	3.54	5.00	14.00	9.00
Total	12	6.33	3.82	1.00	14.00	13.00
GMRT						
4. Level A	16	1.00	0.0	1.00	1.00	0.0
5. Level D	21	2.48	0.51	2.00	3.00	1.00
6. Level F	21	2.48	0.51	2.00	3.00	1.00
Total	58	2.07	0.79	1.00	3.00	2.00
SAT						
7. Primary I	33	1.15	0.44	1.00	3.00	2.00
8. Intermediate I	24	2.50	1.06	1.00	5.00	4.00
9. Intermediate II	25	2.56	1.16	1.00	5.00	4.00
10. High School	13	5.00	2.04	2.00	8.00	6.00
Total	95	2.39	1.65	1.00	8.00	7.00

approximated the relative importance of selections and choices in the tests. For example, if a pupil did not understand a given selection, theoretically he was likely to get the questions with it wrong. Similarly, a pupil did not need to understand each choice in order to get a question right. Generally a pupil needed only to identify the right answer, eliminate the wrong answer, or give an educated guess about the choices in that "set." In any case, only one question was affected.

Results and Discussion

The readability data are presented and discussed here in the form of general conclusions about the four objectives of this study. Selected data tables are included in the text to follow; however, for the reader who is interested in more specific results, tables of unweighted and weighted means, standard deviations, minimums, maximums and ranges of readability and difficulty scores by test are presented in Appendix A, Tables 2-22. Tables of Pearson product-moment correlations among readability and difficulty scores by test are also presented in Appendix A, Tables 23-32.

The first objective of this study was to characterize *the nature of reading comprehension as tested at three grade levels* (grades 1-2, 4-6 and 9-14).

In order to determine the readability characteristics common to the *CAT*, *GMRT*, and *SAT*, the data of the three test batteries were combined for the lowest level tests (tests intended for grades 1-2), for the intermediate level tests (tests intended for grades 4-6) and for the advanced level tests (tests intended for grades 9-14).[1]

Table 6 presents unweighted means and standard deviations for the number of words in the reading selections, questions and choices, as well as the number of non-Spache and non-Dale-Chall words by test level.[2] The minimum, maximum and range for scores on Table 6 are presented on Table 22 in Appendix A.

Eight generalizations applied to the readability scores:

1. The higher the grade level of the test, the longer and harder its reading selections, questions and choices.[3] For example, as seen in Table 6, reading selections on the lowest level tests had a mean length of 18.51 words, the mean length of selections in the intermediate level tests was 64.71 words and the mean length for selections in the advanced level tests was 130.33 words. A similar trend of increases was observed for the other readability scores.

[1] Results of three different standardized tests have been used by the U.S. Office of Education in assessing performance contractors (Klein, 1971, p. 2). The assumption seemed to be that three tests give a more valid appraisal of relevant student performance than one test.

The tests used in this study were combined only when means and standard deviations of the three test batteries showed similar trends. Scores showing contrary trends would tend to cancel each other out and distort interpretation.

[2] Since data for each test item on the CAT, GMRT, and SAT were combined, tests with more selections, questions and choices were "weighted" more than shorter tests. Usually, this resulted in the SAT being "weighted" more than either the CAT or the GMRT. The GMRT was "weighted" more than the CAT, e.g. in the lowest level tests, the CAT had 15 test items, the GMRT had 34 test items and the SAT had 38.

[3] "Hard" refers to the number of non-Spache and non-Dale-Chall words which also determine the Spache and Dale-Chall ratios and grade scores.

It should be noted here that although length of reading matter appeared to discriminate among test levels consistently, it has not as yet been considered a factor in readability formulae.

Table 6

Mean and Standard Deviation of the Number of Words, Non-Spache and Non-Dale-Chall Words in the Reading Selections, Questions and Choices by Test Level[a]

Test		Number of Words		Non-Spache Words		Non-Dale-Chall Words	
Levels	N	Mean	S.D.	Mean	S.D.	Mean	S.D.
Lowest (Grades 1-2)							
Selections	53	18.51	7.59	1.19	1.36	0.13	0.59
Questions[b]	87	6.23	4.16	0.29	0.57	0.01	0.11
Choices [c]	326	0.82	0.87	0.16	0.39	0.03	0.17
Intermediate (Grades 4-6)							
Selections	73	64.71	42.48	16.19	12.50	9.58	8.45
Questions	206	18.13	9.59	4.55	2.69	2.75	2.51
Choices	876	1.34	0.81	0.66	0.53	0.34	0.50
Advanced (Grades 9-14)							
Selections	39	130.33	142.04	41.08	40.29	33.72	35.83
Questions	162	18.85	9.51	6.48	3.63	5.07	2.80
Choices	700	2.09	2.32	1.06	0.75	0.88	0.80

[a] Data for the three test batteries (CAT, GMRT, SAT) are combined

[b] Some questions were implicit in the selection and thus were counted as having "0" words, e.g. His best bubble pipe was on the top shelf (GMRT, Level A, item 17). The sentence was followed by picture choices, one of which was a bubble pipe on a top shelf. The sentence's readability scores were included under those of the selections of the test.

[c] Some of these choices were word choices and others were picture choices. Sometimes picture choices had no words at all. Other times picture choices had words that were unnecessary in determining the correct answer, e.g. "These blocks have letters on them (GMRT, Level A, item 15)." The choices were pictures of 1) cubes without letters, 2) cubes with letters, 3) envelopes with addresses on them, 4) rectangular blocks without letters. It was not necessary in answering the question to be able to read the addresses on the envelopes; thus a readability score was not computed for that picture choice. However, sometimes picture choices had words that were necessary for determining the right answer, e.g. "Which sign means that skating is not allowed (GMRT, Level A, item 25)." The choices were pictures of signs with writing on them 1) Hockey not Allowed, 2) No Skating, 3) Skating Permitted, 4) Racing not Allowed. Since the signs had to be read in determining the correct answer, readability counts were made for each choice.

Consequently, tests at higher levels seemed to require larger and broader vocabularies. They also seemed to expect pupils to assimilate and process a greater amount of information. However, on the basis of the readability analysis it is not possible to characterize the vocabulary or the information pupils were expected to understand.

2. The differences in lengths and number of hard words of reading selections, questions and choices from the lower level to the intermediate level tests were consistently greater than the differences from the intermediate level to the advanced level tests. For example, the mean lengths of questions in tests intended for grades 1-2 was 6.23 words. The mean length of questions in tests intended for grades 4-6 was 18.13 words, and the mean length of questions in tests intended for grades 9-14 was 18.85 words.

Thus, the average question in the intermediate level tests was almost 3 times as long as the average question in the lowest level tests. But the average question in the advanced level tests was almost the same length as the average question in the intermediate level tests. Although magnitudes differed, similar relationships appeared among selection and choice lengths, as well as among hard word scores for the selections, questions and choices.

Size of the increase from test level to test level seemed to depend on the readability score and the part of the item, e.g.

average question length, as noted above, increased 3 times from the lowest level to the intermediate level tests while hardly at all from the intermediate level to the advanced level tests. Average selection length similarly increased about 3 times from the lowest level to the intermediate level tests. But it also increased about 2 times from the intermediate level to the advanced level tests. Furthermore, the average number of non-Spache words in the selection increased by 8.5 times from the lowest to the intermediate level tests and 2.5 times from the intermediate to the advanced level tests.

3. Usually, the higher grade level tests contained selections, questions, and choices with more diverse lengths and more diverse numbers of hard words. For example, the larger the mean of the number of non-Spache words, the larger the standard deviation of the number of non-Spache words. The mean for non-Spache words in the selection of tests intended for grades 1-2 was 1.19 with a standard deviation of 1.36. In tests intended for grades 4-6, the mean number of non-Spache words in the selection was 16.19 and the standard deviation was 12.50. Tests intended for grades 9-14 had a mean number of 41.08 non-Spache words in their selections and a standard deviation of 40.29. Similar increases in standard deviations existed for questions and for the other selection, question and choice readability scores on Table 6.[1]

[1] The one exception that existed was that the number of words in the choices of the lowest level tests had a slightly higher standard deviation than the intermediate level tests. This may have been due in part to the fact that on lowest level tests some choices were pictures without words and other choices were pictures with many words.

Thus, selections, questions and choices of lower level tests had more uniform lengths and numbers of hard words than higher level tests. Although this generalization applied to the number of hard words, it was not always consistent with the ratios of hard words to the number of words in the selections, questions or choices. Table 7 presents ratios of the number of sentences to the number of words in the reading selections (average sentence length), the number of non-Spache words to the number of words in the reading selection (Spache ratio), and the number of non-Dale-Chall words to the number of words in the reading selection (Dale-Chall ratio). Weighted means for the ratios are presented in Table 17, Appendix A. Corresponding ratios for questions and choices are on Tables 19 and 20 in Appendix A. Table 7 also presents grade scores for the two readability formulae.

4. With some exceptions the selections, questions and choices in tests intended for grades 9-14 had more uniform Spache and Dale-Chall ratios than tests intended for grades 4-6. For example, the standard deviation for the Spache ratio of selections in the intermediate level tests was 8.30, while the standard deviation for the Spache ratio of selections in the advanced level tests was 6.18. Thus, the number of words, the number of hard words, and the diversity of these counts were greater in the reading selections, questions and choices of tests intended for higher grade levels than in tests intended for lower grade levels. However, the diversity of hard word ratios (Spache and Dale-Chall ratios) in the selections, questions and choices did not consistently increase.

Table 7

Mean and Standard Deviation for Sentence Length, Spache Ratio, Dale-Chall Ratio, Spache Grade Score and Dale-Chall Grade Score in the Reading Selections by Level[a]

				Ratios				Grade Scores			
	Number of	Sentence Length		Spache		Dale-Chall		Spache		Dale-Chall	
Test Level	Selections	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Lowest (Grades 1-2)	53	6.44	2.30	6.35	7.59	0.64	2.43	2.29	0.88	4.09[b]	0.52
Intermediate (Grades 4-6)	73	16.98	6.29	25.26	8.30	13.99	8.87	5.41[b]	1.10	8.02	2.75
Advanced (Grades 9-14)	39	23.96	8.87	33.70	6.18	26.97	7.97	7.12[b]	1.41	12.91	2.50

[a] Scores for the three test batteries (CAT, GMRT, SAT) are combined. A description of how scores were compiled is given on pp. 46,47.

[b] Invalid as grade scores. As noted above, the Spache formula was intended by its author only for grades 1 to 3. The Dale-Chall formula was intended by its authors only for grades 4 and above. Consequently, the Spache grade scores for the intermediate and advanced test levels and the Dale-Chall grade score for the lowest test level only demonstrate the relationships existing among the test levels and between the readability formulae.

5. The ratios of hard words in the reading selections and questions were more similar to each other than to the ratio of hard words in the choices. For example, the advanced level tests had a mean Spache ratio of 33.70 for the selections and an average Spache ratio of 35.66 for the questions (Table 19, Appendix A). However, the average Spache ratio for the choices was much larger, 67.75 (Table 20, Appendix A). The reason for these similarities and differences may be that while selections and questions were usually made up of sentences that included simple words like articles and conjunctions, choices were usually few isolated words of more uniform difficulty.

No one seems to know what the ratio of hard words in reading selections of tests, school readers, or textbooks should be.

The Dale-Chall ratio indicated the frequency of hard words in a given reading selection, question or choice.[1] Thus, according to the Dale-Chall ratio, reading selections had about 1 hard word in 200 in tests intended for grades 1-2, about 1 hard word in 10 in tests intended for grades 4-6, and about 1 hard word in 4 in tests intended for grades 9-14 (Table 7, p. 61). Questions had a similar progression, though usually higher frequencies, e.g. about 1 hard word in 100 words in lower level tests, about 1 hard word in 6 words in intermediate level tests and about 1 hard word in 4 words on advanced level tests (Table 19,

[1] Non-Spache words do not appear on the International Kindergarten Union List and the first 1000 words of Thorndike's Teachers Word Book (Spache, 1953). These lists probably include most words known by 1st graders. Non-Dale-Chall words do not appear on the Dale List of 3000 Familiar Words. This list includes words known by 80% of a sample of 4th graders. Since 4th graders understand more words than most 1st graders, non-Spache words include more "simple" words than non-Dale-Chall words. Also since 4th grade reading level is the established literacy criterion in this country, words not known by most 4th graders may be viewed as generally difficult.

Appendix A). Choices had a similar progression, but even higher frequencies, e.g. about 1 hard word in 50 on lowest level tests, 1 hard word in 4 on intermediate level tests, and 1 hard word in 2 on advanced level tests.[1]

Before the discussion of the underlying meaning of readability scores, three more findings are presented.

6. Grade scores of the two readability formulae were not identical. As noted above the Spache formula was intended by its author only for grades 1 to 3. The Dale-Chall formula was intended by its authors only for grades 4 and above. Consequently, the Spache grade score for the intermediate and advanced test levels and the Dale-Chall grade score for the lowest test level do not refer to the grade level of children for whom these tests are appropriate. These grade scores were presented only to demonstrate the relationship existing among the test levels and between readability formulae. The remaining scores however may give an indication of appropriate grade level.

The predictions made by readability formulae are generally accurate and reliable within the range of about one year (Chall, 1958b; Dyer, 1971). Thus, the Spache appraisal of the lowest level test (2.29) seemed to correspond to the grades for which the authors intended the test. The Dale-Chall appraisal of the intermediate level test (8.02) seemed higher than suitable for the grades intended. And, the Dale-Chall appraisal of the advanced level test (12.91) appeared to correspond to the grades for which the tests were intended. One reason for differences between readability and test-authors' appraisals

[1] Test-authors generally intend that reading comprehension tests resemble reading matter in school books. In fact, one test-author suggested that a school intending to use his tests, "examine its own curriculum and the test content to ascertain whether or not the latter satisfactorily covers the former (Kelley, et al, 1966, p. 23)."

may be the criterion used in establishing "grade appropriateness." Readability formulae used items passed by 50%-75% of the pupils in a given grade. Test-authors sometimes used items passed by only 26% of the pupils in a given grade. Furthermore, although the mean grade score of items on a given test seemed to correspond to the grades for which the test was intended, little can be noted about the "grade appropriateness" of individual selections.

7. Weighting readability scores by the number of questions had little effect on the relationships among levels. Mostly the direction of relationships remained the same although the sizes of relationships were somewhat increased or decreased. For example, the unweighted means for the number of sentences in the reading selections at the three test levels analyzed were 3.12, 5.88 and 9.51 (Table 3, Appendix A). The weighted means for the same three levels on the same score were 2.36, 6.11, and 10.84 (Table 14, Appendix A).

The greatest proportional effect of weighting scores seemed to be in the reduction of means on the lowest level tests. This may have been due to the inclusion of "0" scores for the test items at that level which had no selections.

8. One interesting side note is the relationship between adjacent level tests. The only adjacent level tests in this study were the SAT Intermediate I, intended for grade 4, and the SAT Intermediate II intended for grades 5 and 6.[1]

[1] All other tests skipped in-between levels. For example, between the CAT Lower Primary and Elementary tests analyzed in this study, is a CAT Upper Primary not analyzed in this study.

The increases in the means of readability scores were generally as consistent in the adjacent tests as in tests where levels were skipped. However, the SAT Intermediate I more frequently had higher minimum and maximum scores than the SAT Intermediate II. For example, the mean number of words in the reading selections of the SAT Intermediate I was 64.79, and of the SAT Intermediate II was 67.04. Although the average reading selection on the SAT Intermediate II was a bit longer, both the shortest and longest selections on this test were shorter than the corresponding selections on the Intermediate I, e.g. the longest reading selection on the Intermediate I had 161 words, and on the Intermediate II had only 127 words. The shortest reading selection on the Intermediate I test had 13 words and the shortest selection on the Intermediate II had 12 words. Such reversals occurred infrequently when in-between test levels were skipped, e.g. the highest Spache ratio for a reading selection in the SAT Intermediate level tests was 50.00, while the highest Spache ratio for a reading selection on the SAT advanced level test was only 39.39 (Table 17, Appendix A).

On the whole, the reading comprehension subtests analyzed consistently increased in the number of words as well as in the number and ratio of hard words in the average reading selections, questions and choices at each higher test level. Generally, differences between levels were not uniform. Greater differences in readability scores existed from the lower to the intermediate than from the intermediate to the advanced level tests.

The significance of these findings rests not only in the scores themselves. Readability scores reflect more fundamental factors about the language used to write reading selections, questions and choices. Horn stated:

> ...difficulty of vocabulary is tied up with the remoteness of the concepts from the reader's experience; and a large number of different words and long involved sentences are related to the complexity of the ideas presented. (Horn, 1937, p. 170)

Therefore, pupils were tested on more "remote concepts" and on more "complex ideas" in the selections, questions and choices of higher level tests. Also, the present analysis suggested that "remoteness" and "complexity" increased more from the lowest level tests to the intermediate level tests, than from the intermediate to the advanced level tests. Further investigation was undertaken to determine whether or not these differences were consistent in different reading comprehension subtests.

The second objective of this study was to characterize the nature of reading comprehension as tested by different test batteries.

The readability scores of the reading comprehension subtests of three widely used test batteries (California Achievement Test, 1963; Gates-MacGinitie Reading Test, 1965, 1969; and Stanford Achievement Test, 1964, 1965) were contrasted. Table 8 presents unweighted means and standard deviations for the number of words in the selections, questions and choices. Minimums, maximums and ranges for these scores can be foundin Tables 2, 6, and 9 in Appendix A.

Table 8

Mean and Standard Deviation of the Number of Words in the Reading Selections, Questions, and Choices by Test

		Selections			Questions			Choices		
		Number	Mean	S.D.	Number	Mean	S.D.	Number	Mean	S.D.
	CAT									
1.	Lower Primary	4	23.00	5.83	15	12.53	3.94	38	1.68	1.51
2.	Elementary	3	198.00	60.22	30	18.23	14.43	120	2.10	1.33
3.	Advanced	5	418.60	161.83	45	13.64	5.70	180	3.55	2.93
	Total	12	231.58	205.47	90	14.99	9.58	338	2.82	2.46
	GMRT									
4.	Level A	16	20.38	8.64	34	4.94	3.81	136	0.35[a]	0.84
5.	Level D	21	42.81	9.77	52	16.38	5.91	260	1.00	0.06
6.	Level F	21	57.33	20.16	52	25.02	10.32	260	1.00	0.0
	Total	58	41.88	20.34	138	16.82	10.80	656	0.87	0.46
	SAT									
7.	Primary I	33	17.06	7.00	38	4.90	1.41	152	1.01	0.11
8.	Intermediate I	24	64.79	38.30	60	18.85	8.81	240	1.31	0.71
9.	Intermediate II	25	67.04	31.29	64	18.81	9.94	256	1.36	0.76
10.	High School	13	137.38	102.34	65	17.52	8.28	260	2.16	2.43
	Total	95	58.74	58.78	227	16.12	9.66	908	1.52	1.47

[a] Many of the choices were pictures without any words. Some choices were pictures with words.

Four generalizations were made on the basis of the analysis by test batteries.

1. Findings about test levels within each battery were similar to the findings about test levels when batteries were combined:

   a. the higher the grade level within each battery, the longer and harder its reading selections, questions and choices.[1]

   b. the differences in lengths and number of hard words in the reading selections, questions and choices from the lowest level to the intermediate level tests were proportionately greater than the differences from the intermediate to the advanced level tests.

   c. the higher grade level tests usually contained selections, questions and choices with more diverse lengths and more diverse number of hard words. More specifically, the higher the means of these scores, the higher their standard deviations.

[1] The following considerations should be noted in applying these generalizations to choices. First, the GMRT, Level A had picture choices with and without words. Second, the range for the number of words in the choices of the 3 test batteries varied. The CAT choices ranged from 1 to 19 words (Table 9, Appendix A). The SAT had a similar range of 1 to 14 words. However, the choices on the GMRT ranged only from 1-3 words, and in Surveys D and F, the choices were almost uniformly 1 word.

Another exception was the number of words in the questions. Both the CAT and GMRT had fewer words in the questions of the advanced level test than in the questions of the intermediate level tests (Table 6, Appendix A).

The data from which these generalizations were made are in Tables 2-22, Appendix A.

d. more often than not, selections, questions and choices in tests intended for grades 9-14 had more uniform Spache and Dale-Chall ratios than tests intended for grades 4-6.

e. the ratios of hard words in the reading selections and questions were more similar to each other, than to the ratio of hard words in the choices.

For example, selection length increased from lower level to higher level tests in each battery, e.g. on the CAT the average length for reading selections was 23 words at the lowest level test, 198 words at the intermediate level test and 418 words at the advanced level test. Similarly, on the GMRT, reading selections had an average length of 20.38 words at the lowest test level, 42.81 words at the intermediate level and 57.33 words at the advanced level. In keeping with this pattern, the SAT reading selections had an average of 17.06 words at the lowest test level, 65.91 words at the intermediate level and 137.38 words at the advanced level. Generally, question and choice lengths increased from level to level in each test battery as well. Other readability scores usually also increased with higher test level.

However, as seen from these data, the average length for reading selections at a given level was not the same in the three test batteries analyzed; nor, generally, were the increases from level to level the same.

2. The CAT, GMRT and SAT usually differed in the average number of

words and number of hard words they contained in reading selections, questions and choices at a given grade level. For example, on the advanced level tests reading selections in the CAT were longest with an average length of 418 words. Reading selections in the GMRT were shortest with an average length of 57.33 words. Reading selections in the SAT had an in-between average length of 137.38 words.

Among the 3 test batteries analyzed, the GMRT most often had the longest questions with the most hard words. For example, on the advanced level tests, the questions in the CAT were shortest with an average of 13.64 words.[1] The questions in the GMRT were longest with an average of 25.02 words. And, the questions in the SAT were in-between with an average of 17.52 words.

In part, this may have been due to the fact that questions on the different tests often took different forms. The CAT always had separate questions in either sentence completion or direct question form. The GMRT, Primary A had either regular questions or questions implicit in the selection, e.g. a direction telling a pupil to mark one of the choices, or a description of one of the picture choices which the pupil was expected to mark. In Surveys D and F, the GMRT always had cloze-like blanks in the reading selections. The SAT had questions in all these forms.

[1] The CAT Lower Primary and Elementary were exceptions since instructions on how to respond immediately preceded the questions. When such instructions were not read out loud by the teacher and were necessary for getting the question right, they were added to the readability counts of the question. The questions were thereby artificially lengthened at these levels.

3. The increases in readability scores from test level to next higher test level differed for the CAT, GMRT and SAT. For example, the reading selections of the intermediate level CAT were over 8 times as long as those on the lowest level CAT. The intermediate level GMRT had reading selections about twice as long as the lowest level GMRT. And intermediate level SAT reading selections were about 4 times as long as lowest level SAT reading selections. Similar differences in increases existed among the other readability counts of selections, questions and choices (Tables 3-20, Appendix A).

Despite these differences in readability counts among tests, some similarities were observed in ratios.

4. The average sentence length in the selections as well as ratios of hard words to total words in the selections and questions were relatively uniform among tests. For example, on the intermediate level questions, the CAT had a Spache ratio of 24.76, the GMRT had a Spache ratio of 26.89 and the SAT had an average Spache ratio of 25.94 (Table 19, Appendix A).

This was not the case with choices. For instance, on the advanced level test the CAT had a Spache ratio of 50.97, the GMRT had a Spache ratio of 84.99 and the SAT had a Spache ratio of 67.30.

Interesting patterns emerged from the relationships found among batteries. For example, Figure 1 demonstrates the ranks of the CAT, GMRT, and SAT on the average number of words in the reading selections, questions and choices.[1]

[1] All levels analyzed within one test battery, e.g. CAT, were combined. Although ranking the scores tends to magnify the differences, it clarifies the general patterns.

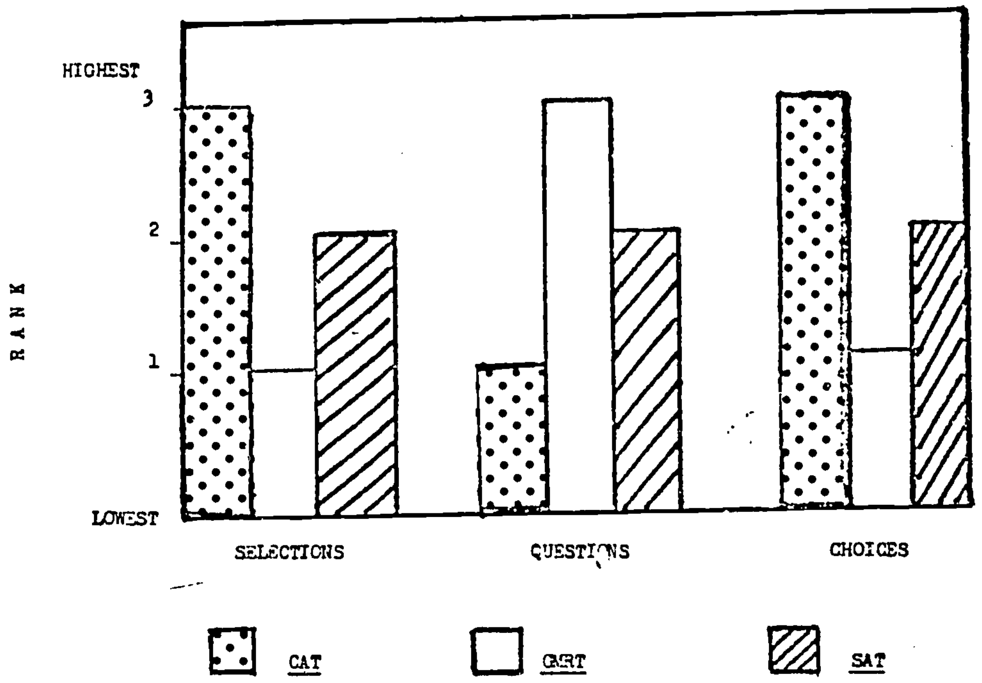

Figure 1: Ranks of three reading comprehension tests on the number of words in their items.

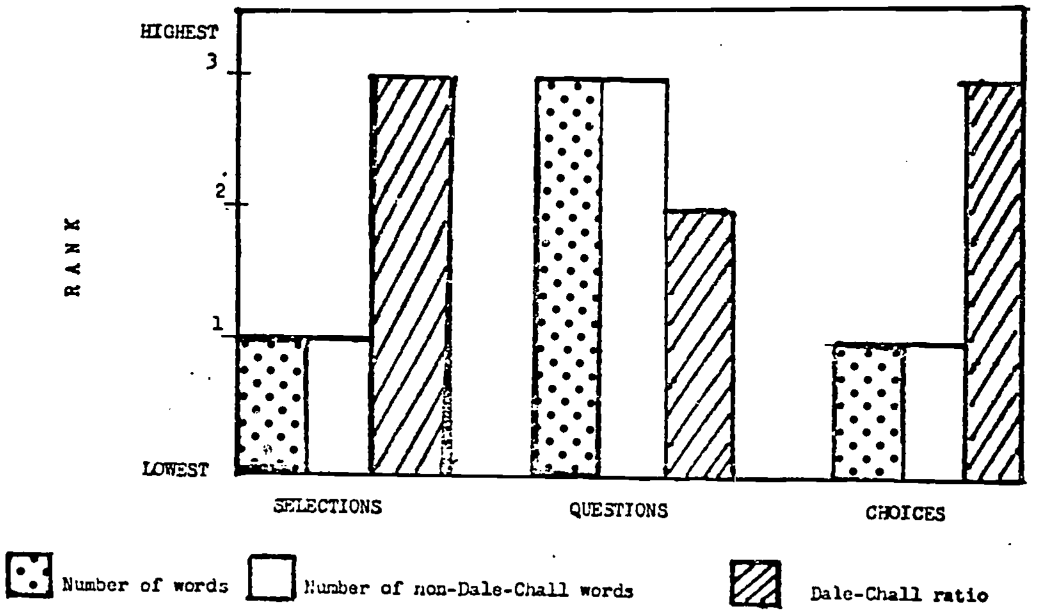

Figure 2: Rank of the Advanced level GMRT, relative to the Advanced level CAT and SAT, on three readability scores: number of words, number of non-Dale-Chall words, Dale-Chall ratio.

Characteristically, the *CAT* had longer selections and choices but shorter questions than the *GMRT* or the *SAT*. Characteristically, the *GMRT* had shorter reading selections and choices, but longer questions than the *CAT* or the *SAT*. The *SAT* scores seemed consistently in-between those of the *CAT* and the *GMRT*. More often than not, the rank for the levels combined accurately reflected the rank at any given level. For example, at the advanced level, the *CAT* had longer reading selections and choices, but shorter questions than the *GMRT* or *SAT*. The *GMRT* at that level had the opposite, i.e. shorter selection and choices, but longer questions than the *CAT* or *SAT*. The *SAT* again remained between the *CAT* and *GMRT*.

Although similar patterns emerged for other readability counts, usually opposite patterns occurred for readability ratios, i.e. when the rank of readability counts in a test went up, the rank of the ratios in that test went down and *vice versa*.[1] For example, Figure 2 presents the rank of the advanced level *GMRT*, in comparison to the advanced level *CAT* and *SAT*, on the number of words, the number of non-Dale-Chall words and the Dale-Chall ratios in the reading selections, questions and choices.

Of the subtests analyzed at the advanced level, the *GMRT* had the fewest words, the fewest hard words (non-Dale-Chall words) but the highest proportion of hard words (Dale-Chall ratio) in the selections and choices. The reverse was true for questions, i.e. questions had

[1] Inconsistencies in the patterns occurred among the ranks of the tests intended for grades 1-2 and among question readability scores.

the most words, the most hard words but a smaller proportion of hard words. Similar patterns emerged for other test levels.

As has been shown, the readability scores of the CAT, GMRT, and SAT differed. The differences tended to fall into characteristic patterns. Tests with longer selections and choices had shorter questions. Tests with more words and more hard words in their selections, questions or choices had a lower proportion of hard words.

The third objective of this study was to identify factors that may contribute to difficulty in tested comprehension.

The empirical criteria of the difficulty of test items were the difficulty scores provided by test publishers. These difficulty scores represented the percentage of pupils who answered a given item correctly (see p. 49). Means, standard deviations, minimums, maximums and ranges of difficulty scores by grade and test are presented in Table 21, Appendix A.

The distribution of difficulty scores led to the following 4 observations:

1. Generally the minimum and maximum difficulty scores increased for each higher grade included in a given test level. For example, the intermediate level CAT was intended for grades 4, 5, and 6. In the 4th grade the minimum percentage of pupils passing a given test item was 12.3; while the maximum percentage was 81.5. In the 5th grade the minimum difficulty score rose to 18.5 and the maximum rose to 92.4.

2. Although increases occurred generally, both the size of the minimum and maximum difficulty scores and the size of the increases were

different for the CAT, GMRT and SAT. For example, in the GMRT the 4th grade difficulty score minimum was 2.2 and the maximum was 92.0. In the 6th grade GMRT the difficulty score minimum was 22.0 while the maximum was 95.7. Thus, the difference in minimum difficulty scores from the 4th to the 6th grade CAT was 11.5 (see CAT scores in "1" above), while the difference in minimum scores from the 4th to the 6th grade GMRT was 19.8.

3. The CAT, GMRT and SAT had different ranges of difficulty scores. The CAT generally had the narrowest ranges of difficulty scores -- the smallest range was 46.1 and the widest range in that battery was 69.2. The GMRT generally had the widest ranges of difficulty scores. The smallest GMRT range was 73.7 and the widest range was 89.8. The SAT had in-between ranges of difficulty scores; 60.0 was the smallest in that battery and 77.0 was the widest.

4. The CAT, GMRT and SAT also had different means for difficulty scores. Mean difficulty scores on the CAT went from 26.0 to 64.2, on the GMRT from 48.5 to 71.7 and on the SAT from 43.3 to 61.3. Generally then, the CAT had the lowest means, the GMRT had the highest means and the SAT had means that fell in-between.

The differences among difficulty scores reflected the differences in the criteria used by the constructors of the three test batteries. For example, while the designers of the CAT chose comprehension test items that were passed by an average of 26% to 64% of the pupils at a given grade, the designers of the GMRT seemed to prefer comprehension test items passed by a greater percentage of pupils--an average of

48.5% to 71.7% of pupils at a given grade. The differences among difficulty scores may also reflect the particular sample of pupils who took the tests. Different groups of pupils are likely to produce different difficulty scores.

Due to these differences, difficulty scores were studied only by individual tests (for correlations of readability and difficulty scores by test see Tables 23-32 in Appendix A). Four generalizations were made on the basis of the correlation analysis.

1. The highest correlations existed consistently among the difficulty scores themselves. Hence, it seemed that factors which made an item difficult at one grade level tended to be very closely related to factors which made the same item difficult at another grade level. For example, in the CAT intermediate level test (Table 24, Appendix A), which is intended for grades 4, 5 and 6, the correlation coefficient for the difficulty scores of grade 4 and 5 was .98. The correlation coefficient for difficulty scores of grade 5 and 6 was .95. Correlations of difficulty and readability scores were lower. The highest correlation coefficient for a difficulty and a readability score was .91--the correlation of the items' difficulty scores in grade 4 and the number of non-Dale-Chall words in the selections.

2. Although the difficulty scores for different grades were consistently and highly intercorrelated, the difficulty scores for one grade correlated quite differently with readability scores than the difficulty scores for the other grades. For example, on the lowest level CAT (Table 23, Appendix A) the correlation coefficients of difficulty and

readability scores for the 1st grade were often more than twice as large as the correlation coefficients for the 2nd grade, e.g. 1st grade difficulty scores and number of words in the selections had a correlation coefficient of -.67. Second grade difficulty scores and the number of words in the selections had a correlation coefficient of -.29.

3. Difficulty scores correlated "most highly" with different readability scores in the 10 subtests analyzed. For example, on the lowest CAT (Table 23, Appendix A), difficulty scores correlated most highly with the following readability scores: Dale-Chall grade score for the selection (-.71), the number of sentences in the selection (-.71) and the number of words in the question (-.70). At the intermediate level, the CAT difficulty scores (Table 24, Appendix A) correlated most highly with different readability scores: number of non-Dale-Chall words in the selection (-.91, -.88, -.88), and the number of non-Spache words in the selection (-.90, -.87, -.87).[1] Furthermore, the GMRT's difficulty scores at the intermediate level (Table 27, Appendix A) correlated most highly with another combination of readability scores: Dale-Chall grade score for the selection (-.60 and -.62), Dale-Chall ratio in the selections (-.55 and -.56), questions (-.51 to -.55), and choices (-.53 and -.54).[2]

[1] The number of correlation coefficients corresponds to the number of grades for which difficulty scores were available, i.e. grades 4,5 and 6.

[2] The correlation coefficients correspond to correlations of the readability scores with difficulty scores for grades 4 and 6.

Correlation coefficients demonstrate the relationship between two groups of scores. It has been shown that difficulty scores do not consistently correlate with the same readability scores in the different subtests analyzed. In order to find the effect these readability scores have in determining difficulty scores in a given test, either many partial correlations would have to be computed or a multiple regression analysis would have to be conducted.

4. Generally the difficulty scores had higher correlations with the selections' readability scores than with the readability scores of either questions or choices. For example, correlation coefficients of the CAT advanced level test (Table 25, Appendix A), ranged from .04 to -.55 for difficulty scores and selection readability scores; from .02 to -.21 for difficulty scores and question readability scores; and from .01 to .20 for difficulty scores and choice readability scores. Similar relationships appeared in other tests as well.

In order to find the effect that selection, question or choice readability scores have in determining difficulty scores, a multiple regression analysis would have to be conducted. However, generally, the correlation coefficients between readability and difficulty scores were not very high. For example, only the CAT elementary and intermediate level subtests had correlation coefficients for readability and difficulty scores that were over .70. Usually these coefficients were much lower (see Tables 23-32, Appendix A). Therefore, it appears that factors other than "readability" also influence item difficulty.

The fourth objective of this study was to characterize the nature of tested reading comprehension.

On the basis of the readability analysis, four conclusions about tested reading comprehension seemed appropriate:

1. Comprehension was usually tested by longer selections, questions, and choices at more advanced levels. Also at more advanced levels, the selections, questions and choices contained more hard words and a greater ratio of hard words to the number of words. Furthermore, the increases in readability scores from level to level were not uniform. Greater increases appeared between lower levels than between higher levels on almost all scores.

2. Comprehension appeared to be tested somewhat differently by the CAT, GMRT and SAT. While one test battery had more words, more hard words and relatively small hard words/number of words ratio in its selections and choices, another battery had fewer words, fewer hard words and a relatively high hard words/number of words ratio in its selections and choices. The subtests at the 3 levels analyzed within each battery however, seemed to be relatively consistent.

3. Empirical difficulty of comprehension test items seemed to be "most correlated" with a different set of readability scores in each subtest. Usually item difficulty was more related to readability scores of selections than readability scores of questions or choices.

4. Empirical difficulty of tested reading comprehension did not seem very closely related to readability factors in general. While on some subtests, item difficulty and readability scores were highly correlated,

on most subtests, the correlations were not very high. This finding suggested that other factors may heavily influence the difficulty of tested reading comprehension.

The readability analysis presented thus far has outlined "stylistic elements" of reading comprrhension tests. The fact that selections, questions and choices get longer and have more words suggested that the ideas presented became more complex and unfamiliar. However, what the nature of the complexity and unfamiliarity was has not been revealed.[1] In the section to follow the content of the subject matter as well as the tasks to be performed on reading comprehension subtests were studied.

[1] Chall (1958b,p. 156) stated that according to the "judgment of experts (teachers, librarians, and publishers) and readers...content, stylistic elements, format, and organization contribute to difficulty. Stylistic elements are represented by the readability analysis. Some description of the format of tests and test items has also been included in this chapter.

CHAPTER V

Task Measurement

Introduction

Tasks for comprehension are empirically established by reading comprehension test items. Successfully performing the task means answering the test item correctly. The knowledge and behavior necessary to perform well on reading comprehension tests has not as yet been clearly specified.

Wittgenstein suggested that we "think of words as instruments characterized by their use (in Chihara and Fodor, 1966, p. 388)." By use he meant, according to Pitcher (1964), saying or writing the word, following directions involving the word, fetching or drawing what the word represented and also discriminating the object the word represented from other objects. According to Wittgenstein, accurately using words in these ways demonstrates one's understanding.

Although Wittgenstein's concern was with the meanings of words, the same strategies for demonstrating understanding may also apply to larger language units such as sentences, paragraphs and stories. In fact the strategies he suggested resembled the questions developed by Thorndike that still are used in modified form on current standardized reading comprehension tests. For example, on several items in the lowest level standardized comprehension subtests analyzed, pupils are required to mark the picture that corresponds to a word in a direction, e.g. "Mark the cat (GMRT, Level A, item 1)." The direction is accompanied by four picture choices 1) a cat in front of a fireplace 2) a chair 3) a chair in front of a fireplace 5) a dog. Such items require the pupil to follow directions that involve the object and to perform the behavior described by the words, as well as to discriminate the object from other objects. The more advanced comprehension tests analyzed required the pupil to discriminate the correct answer from among a number of possible answers (multiple-choice distractors).

Thorndike indicated that the comprehension task as he defined it was determined by aspects of the reading selection, the question and the responses. Responses on nearly all comprehension test items analyzed are now restricted to discriminating among multiple-choice answers. Therefore, in analyzing the tasks, this study investigated qualitative aspects of the reading selections, the questions and the multiple-choice answers.

A rating scale for reading selections, a rating scale for questions and a rating scale for choices were devised to categorize the item tasks. Much effort was put into specifying sufficient criteria for the ratings so that raters could easily agree,

The rating scales will be presented in the procedure section to follow but a brief description of them is presented here. The reading selections were rated according to topic or content. Eight topics were defined and included in the selection rating scale:

1) riddle 2) story 3) language 4) math 5) social studies 6) social science 7) science 8) humanities.[1] A "0" category was also included for those test items that had no reading selections.

The scale for questions rated the relationship between the way information was presented in the reading selection and the way the same information was presented in the correct answer. Nine categories were defined and included in the question rating scale: 1) recognition 2) contextual paraphrase 3) grammatical paraphrase 4) semantic paraphrase 5) definite concepts 6) probable concepts 7) language concepts 8) previous knowledge 9) word-picture matching. When appropriate, the last category was combined with one of the preceding categories. For example, if the picture in the choice was not clearly described in the selection or question, but was probably the best choice on the basis of indirect clues given, the item was rated as"picture-matching" and "probable concept."

Lastly, the scale for choices rated the relationship of distractors to the selection, to the question and to the correct answer. In all, 12 ratings were defined on the scale: 1) other 2) grammatical 3) associative 4) associative-grammatical 5) categorical 6) categorical-grammatical 7) textual 8) textual-grammatical 9) textual-associative 10) textual-associative-grammatical 11) textual-categorical 12) textual-catégorical-grammatical. Generally, "grammatical" referred to whether or not a given distractor was a grammatical answer to the question. "Associative" referred to a slight semantic or conceptual relationship

[1] The categories for selection, question and choice rating scales were identified during a preliminary study of a number of reading comprehension subtests at different levels by Auerbach (1970). Generally the categories included in the scales reflected the variety of items on reading comprehension tests.

of a given distractor to the correct answer choice, i.e. the distractor represented a feature or function of the word represented by the correct choice. "Categorical" referred to a closer semantic or conceptual relationship of a given distractor to the correct answer choice than "associative", i.e. the "categorical" distractor was the same kind of object as the correct answer choice. "Textual" simply meant that the distractor was a word in the reading selection of the test item. And, "other" meant that the distractor was none of the above; essentially the distractor was irrelevant to the item.

As with readability scores, the significance of these rating scales is not immediately apparent. The scale categories tend to reflect more fundamental factors about the test items. Each scale reflects to some extent, contextual, grammatical, semantic and conceptual aspects of the items, e.g. the scale for questions rates whether or not the information is given explicitly, implicitly or not at all in the context, whether or not grammatical changes have occurred in the words presenting the information, whether or not semantic changes have occurred in the words presenting the information and whether or not the concepts involved are general or academic.

An exploration of reading comprehension tasks follows. Comparisons were made among the three test levels and the three test batteries.

Procedure

Each reading selection, question and distractor in the CAT, GMRT and SAT was coded independently according to the following rating scales:

RATING SCALE FOR SELECTIONS

Code and Definition

0 = No selection[a]

1 = Riddle selection - the selection is a description or clue given to help the pupil identify a common object, act, etc. e.g. "I play with my new toy. It is a 1) ball 2) something 3) little 4) play (Stanford Achievement Test, Primary 1, Form X, Paragraph Meaning, 1)."

2 = Story selection - the selection is about relatively common occurrences, events, people; not academically oriented.

3 = Language selection - the selection is primarily about language usage or literature.

4 = Math selection - the selection is about mathematics or requires mathematical concepts.

5 = Social studies selection - the selection is about history, geography, etc.

6 = Social science selection - the selection is about psychology, sociology, anthropology, etc.

7 = Science selection - the selection is about general science, chemistry, biology, etc.

8 = Humanities selection - the selection is about philosophy, art, theology, etc.

[a] Some items have only questions and choices and no selections, e.g. "Which is the big tree (Gates-MacGinitie Reading Test, Primary A, Form 1, Comprehension 2)?" The question is followed by four choices, one of which is a big tree.

RATING SCALE FOR QUESTIONS

Code and Definition

1 = Recognition:

Choosing the right answer requires recognizing an identical word that appears in the selection in the same general context.

ex: The hedgehog of the Old World is a small mammal similar to a porcupine. When it is in danger it rolls itself up into a ball so that it resembles a pincushion and is protected by its sharp quills.

The Old World hedgehog is a

porcupine   pincushion   plant   mammal

(Gates-MacGinitie Reading Test, Primary C, Form 2, Comprehension 17A)

2 = Contextual paraphrase

Choosing the right answer requires recognizing an identical word that appears in the selection in a different linguistic context.

ex: In the tropics, bacteria grow so rapidly that they quickly destroy rotting plant matter, called humus, in the soil. Tropical soils have little

iron   humus   soil   growth

(Gates-MacGinitie Reading Test, Primary E, Form 3, Comprehension 31)

3 = Grammatical paraphrase

Choosing the right answer requires recognizing a grammatical variant (different number, voice, tense, etc.) of a word that appears in the selection in a different linguistic context.

ex: We all inspire and expire when we breathe. Inspiration is the act of taking into ourselves something which is not a part of us. __________is the act of giving back what we have thus obtained....

1. Expire
2. Inspire
3. Expiration
4. Inspiration

(Stanford Achievement Test, Advanced, Form W, Paragraph Meaning 29)

Note: The correct answer is underlined.

4 = Semantic paraphrase

Choosing the right answer requires recognizing a semantical variant (synonym, translation, paraphrase, etc.) of a word or phrase that appears in the selection in a different linguistic context.

ex: If you look at your hands closely, you will see that the skin has little ridges. The pattern of the ridges on the tip of one of your fingers never changes while you live, and this design is different from that on any other finger in the world. This is why the police can use ________ as a means of identification.

1. photographs	3. handshakes
2. handwriting	4. fingerprints

(Stanford Achievement Test, Intermediate 2, Form W, Paragraph Meaning 4)

5 = Definite concepts

Choosing the right answer requires identifying a "common" concept that

a. is not stated in the selection
b. definitely applies to the instances or attributes mentioned in the selection
c. and is the only choice that meets the above conditions.

ex: The third-grade class went on a trip. They saw the fenced fields, the tall silo, and the powerful tractor. They watched the horses and cows and fed the chicks. They were even allowed to hold the baby rabbits.

They saw many

engines   pigs   trees   animals

(Gates-MacGinitie Reading Test, Primary C, Form 2, Comprehension 1B)

6 = Probable concepts

Choosing the right answer requires identifying a "common" concept that

a. is not stated in the selection
b. applies with a certain degree of appropriateness to the set of attributes or instances mentioned in the selection
c. and is the choice that best meets the above conditions.

ex: (read the selection under 5 above)

The children went to a

farm   zoo   park   circus

(Gates-MacGinitie Reading Test, Primary C, Form 2, Comprehension 1A)

7 = Language concepts

Choosing the right answer depends upon semantic and/or syntactic constructions such as: cliches, collo uialisms, antonyms, relatives, antecedents, etc. which are not stated in the selection, but are suggested by the general theme and/or contextual implications of the selection.

ex: When Jane went shopping for a dress, she bought the least expensive one __________ h:r limited budget.

1. in spite of
2. notwithstanding
3. regardless of
4. on account of

(Stanford Achievement Test, Intermediate II, Form W, Paragraph Meaning 16)

8 = Previous knowledge

Choosing the right answer requires previous knowledge, usually obtained in a formal setting, of specific facts such as dates, names, relationships, places, etc.

ex: From 1850 to 1880, Virgini City held a prominent place in the history of silver and gold mining. Its fabulous production of silver and gold has left a tremendous impression on all who ever heard of it. This production played an important role in financing the Union during the ____________________.

1. War between the States
2. Revolutionary War
3. War of 1812
4. Mexican War

(Stanford Achievement Test, Intermediate II, Form Y, Paragraph Meaning 2)

9 = Word-picture matching

Choosing the right answers requires matching a word to its corresponding picture. If the picture is not a clear and simple representation of the word, other of the above categories may be added. For example, if the picture represents a probable concept, it would be rated "96."

RATING SCALE FOR CHOICES

Definitions

Textual - the distractor is stated in the selection (possibly in a different number, tense, etc. If there are a number of words, the distractor is rated textual when:

a. some of the words are stated explicitly in the selection and some are paraphrased
b. most of the content words are stated explicitly in the selection

Grammatical - the distractor fits the grammatical context of the question. Lexical constraints on this category include:

a verb that can only have an animate subject or object; an adjective that can only modify animate nouns, etc.

Categorical - the distractor fits the same general category of descriptors, objects, events, etc. as the correct choice. This category is determined Ly the word meaning as well as its context in the question, and selection. Where appropriate this refers to distractors that are coordinates, synonymous or antonyms of the correct choice.

Associative - the distractor has "associative value" to either the general theme of the selection or the meaning of the correct choice. This category is not as close to the meaning of the right choice as "categorical" above, yet it is not irrelevant. Where appropriate this refers to distractors that are superordinate, subordinate, functions or features of the correct answer.

Other - the distractor is irrelevant and thus unrelated to either the general theme of the selection or the meaning of the correct choice. It is not found in the reading selection, nor would it be a grammatical answer to the question.

Codes[a]

1 = Other
2 = Grammatical
3 = Associative
4 = Associative-Grammatical
5 = Categorical
6 = Categorical-Grammatical
7 = Textual
8 = Textual-Grammatical
9 = Textual-Associative
10 = Textual-Associative-Grammatical
11 = Textual-Categorical
12 = Textual-Categorical-Grammatical

[a] When a category is not included in the code, it does not apply, e.g. code "2" indicates that the distractor is a grammatical answer to the question, but is not "associative," "categorical," or "textual." Code "7" indicates that the distractor is a word in the text but is not "grammatical," "categorical," or "associative."

Examples

1

First Mother measured the milk, baking powder, shortening, flour, and sugar. Then she mixed these together with two beaten eggs. Finally she poured the batter into a pan and put the pan into the oven.

A. Mother was making

a cake   a dress (2)   cookies (6)[a]   flour (10)[a]

B. She did not use any

milk (8)   salted (5)[a]   pepper   baking (9)[a]   sugar (12)

4

Ruth was busily getting her costume ready for the party. She had already made a tall pointed hat out of black paper. She and her mother had just finished a long black cape. The broom that she would ride was standing in the corner.

A. Ruth was going to the party as a

witch   ghost (6)   cowgirl (4)   costume (11)[a]   pumpkin (4)

B. Ruth still needed a

tall (7)[a]   fun (3)[a]   see (1)[a]   mask

(Gates-MacGinitie Reading Test,
Primary C, Form 1 Comprehension)

Note 1: Numbers in parentheses are example codes.
Note 2: Underlined words are the correct answers.

[a] These distractors were not in the original items. They were included here for the purpose of demonstrating a particular scale category, e.g. in the first paragraph, question A, the choice "cookies" was added to demonstrate a "categorical-grammatical distractor. "Cookies" are not mentioned in the selection yet they are baked goods and essentially the same type of object as a cake. Also, "cookies" completes the question sentence in a grammatically acceptable manner. Another example is the next choice -- "flour." "Flour" is a textual-associative-grammatical distractor. The word "flour" is stated in the text. Flour is an ingredient of a cake and is thus "associative." "Flour" also completes the question sentence in a grammatically acceptable manner.

Some of the original distractors in these items were omitted because they duplicated ratings already demonstrated.

The raters, a Roxbury Latin School senior, a Radcliffe College senior and a Harvard doctoral student were trained as follows:

1. The rater studied the rating scales which presented short definitions for each category, code numbers for each category, and usually example rating for each category.
2. The rater and author discussed the rating scales until the rater stated that he understood them, e.g. the rater asked questions about the scale and checked word definitions.
3. The rater applied the scales to a few random items taken from three test levels but from different forms of the test batteries used in this study.
4. The rater was asked to justify each of his ratings.[1]
5. The rater then applied the scales to the SAT Intermediate I. He again had to justify each rating.

The reading comprehension subtests analyzed in this study were presented to each rater in a different random order. Different random orders were used to avoid biases in ratings that may have resulted from a standard sequence, i.e. coding all the lower level tests in a series,

[1] The rater was asked why he chose a code. He generally replied by referring to part of the definition. Occasionally during the justifying procedure, when raters looked back at the scales and the test items, they spontaneously changed their rating. Where definitions on the scales were not sufficiently clear, they were revised at this point.

or all the tests in one battery in a series might have led raters to using codes that appeared frequently in a mechanical way. By randomly distributing the tests the possibility of raters using the same codes habitually was probably reduced.

When a subtest had been coded by each rater, the results were compared. When differences occurred each rater gave the justification for his code. The justifications were discussed.[1] Generally, a consensus was quickly reached among the raters. The code that all raters agreed on for a given selection, question or choice was the one noted. In the case of 2 questions and 6 choices no consensus was

[1] Sometimes a dissenting rater changed his rating spontaneously after rechecking the category definitions. Other times a dictionary was used to justify ratings, e.g. in choices, to establish whether a given distractor was a synonym, antonym, or feature of the correct answer. Such information determined whether the distractor was coded "associative" or "categorical." Another means of reconciling differences was for each rater to present his reasoning, and also to evaluate the reasoning of the others, e.g. again in choices, the correct answer choice was "flowers," and one of the distractors was "things." One rater contended that "things" was too general and was not really "associative." Another rater reasoned that "things" could be used as a substitute for "flowers" within the context of the reading selection without significantly changing the meaning (see cloze-like item, SAT, Primary 1, 28). the third rater stated that "things" was general but was still relevant and thus should not be coded "other." All raters agreed to rate "things" as "associative." One other approach to reconciling differences was compromise, e.g. one rater coded a distractor as "other," another rater coded the same distractor as "categorical." After trying most of the above approaches, if the raters still could not unanimously agree on one of the given ratings, they compromised at "associative."

reached. One rater contended that his code was as justified as the other.[1] In these 8 cases the code agreed on by two of the raters was used for the analysis.

Treatment of Data

The frequency with which each selection, question and choice scale category appeared on each of the 10 reading comprehension subtests studied was tabulated. The frequency with which the scale categories appeared at each of the 3 test levels and in each of the 3 test batteries was also tabulated.

In addition a comparative study was made of the similarities of items in the reading comprehension subtest of one test battery, and items in other subtests, e.g. word knowledge, science, social studies, in the same battery.

[1] Ratings of choices presented the most problems. The greatest disagreement among raters was in the "associative" and "categorical" codes. On the lower level tests the definition criteria of subordinate, superordinate, coordinate, etc. were applied and fewer disagreements existed. However on the higher level tests when word meanings became more abstract and unfamiliar the judgments became more subjective and the differences among raters more numerous.

Ratings of questions became difficult when two catetories overlapped, i.e. a given question seemed equally appropriate for two categories. For example, a question sentence very closely approximated the selection sentence in which the information was originally given. However, the question sentence was not really identical to the selection sentence in that a modifier was added or omitted or the selection sentence was active while the question sentence was passive. These differences had to be subjectively evaluated and thus one rater coded the question "recognition" while the other coded it "contextual paraphrase."

Ratings of selections presented the fewest problems since they were generally self-explanatory and mutually exclusive.

Results and Discussion

The task data are presented and discussed in the form of general conclusions about the four objectives of this study. Selected data tables are included in the text to follow; however, for the reader who is interested in more specific results, the percentage of each rating in tests and levels is presented in Appendix A, Tables 32-38.

The first objective of this study was to characterize the nature of reading comprehension as tested at three grade levels.

In order to determine the tasks common to the CAT, GMRT and SAT, the data of these test batteries were combined for the lowest level tests, for the intermediate level tests, and for the advanced level tests.

Figure 3 presents the composition of typical items in tests intended for grades 1-2, in tests intended for grades 4-6 and in tests intended for grades 9-14.[1] Tables 33 to 35 in Appendix A present the percentage of reading selections, questions and choices in each scale category at each of the three test levels. Seven generalizations were made on the basis of the task data.

1. Typical reading selections were different on the lowest, intermediate and advanced test levels. On lowest level tests most of the reading selections (71%) were stories. Stories were generally

[1] Typical as used here refers to the most frequently occuring category.

TEST LEVEL	SELECTION	QUESTION	CHOICES
LOWEST (Grades 1-2)	Story	Probable Concept Matching	Associative-Grammatical
INTERMEDIATE (Grades 4-6)	Science	Contextual Paraphrase	Associative-Grammatical Grammatical
ADVANCED (Grades 9-14)	Science Humanities Social Science	Semantic Paraphrase Contextual Paraphrase Probable Concept Previous Knowledge	Associative-Grammatical Grammatical

Figure 3: Typical reading comprehension items at three test levels

about common objects, experiences or people. At the intermediate level reading selections about science were most prevalent (49%). Science selections included general science, biology, physics, etc. Although science selections were also quite frequent in the advanced level tests (31%), selections about humanities (23%) such as art and theology, and about social science (20%), such as psychology, were also numerous. This suggests the second generalization.

2. The range of selection topics became broader at higher test levels. In tests intended for grades 1-2, nearly all the selections (71%) were stories, the next highest category was riddles (24%). There were also 4% science selections at the lowest test level.

Although science selections were most prevalent (49%) in the intermediate test level, 19% of the reading selections were about social studies and 16% of the selections were stories.

As can be seen in Table 33, Appendix A, the reading selections at the advanced level were distributed among even more categories.

3. Reading selections were not only about more topics at each higher test level; they were about more academic topics. Stories about common people, experiences, and events consistently decreased at each higher test level, i.e. 71% of the reading selections at the lowest test level were stories; 16% of the reading selections at the intermediate test level were stories; and, only 5% at the advanced test level were stories. Reading selections about more basic school subjects such as science and social studies, hardly appeared at the lowest test level, were most prevalent at the intermediate level and became fewer at the

advanced level. Reading selections about the more academic subjects, such as social science and humanities, appeared more often in the advanced level tests.

All "school subjects" were not equally represented however. Few reading selections about math, literature and music appeared in comparison to many reading selections about science and social studies.

Consequently, reading comprehension was tested on selected and more "advanced school subjects" at each higher test level. Reading selections resembling reading matter from "life outside of school" were extremely infrequent, especially at the higher test levels. Yet, it would seem that for the greater population, especially those not pursuing academic careers, evaluation on more "everyday" reading matter would be considerably more important than evaluation on academic reading matter. "Everyday" reading matter includes the things a person should be able to read in order to function effectively in today's world, e.g. newspaper articles, advertisements, guarantees, warranties, proposed legislation, trade manuals, job applications, tax forms, instructions for using appliances or tools, directions for cooking or baking, food ingredients, and so on.

The reading of selections represented only one part of the task required by reading comprehension subtests. Another part of the task was using presented information to answer questions correctly.

4. Typical questions were different for the lowest, intermediate and advanced test levels. In the tests intended for grades 1-2, pupils

were asked to identify the words for common objects or generally familiar concepts suggested in the reading selection. Such questions were called "probable concept" (for definitions and examples of question categories see the Rating Scale for Questions, p. 86) and made up 37% of all the questions in tests intended for grades 1-2. Many (26%) of the other questions at the lowest test level asked pupils to match words with corresponding pictures.

Typical test questions at the intermediate level (35%) asked pupils to identify one of the words used in the reading selection as the correct answer to the question. However, the context of the word in the selection was different from the context of the word in the question. Such questions were called "contextual paraphrase."[1]

The typical questions in the advanced level tests were of four types. Twenty-four percent of the questions were "semantic paraphrase." Twenty-one percent of the question at the advanced level were "contextual paraphrase." Eighteen percent of the questions were "probable concepts" and another 18% were "previous knowledge."

The progression of questions from one test level to the next higher test level analyzed seemed to be of two sorts. First, lower level test

[1] The differences between a word's context in the selection and the same word's context in the question varied. Sometimes a logical relationship was established for the two contexts, by the test-author, in the reading selection; at other times it was not. In cloze-like items the word sometimes appeared in the reading selection before the blank (which represented the question), and sometimes after. The effects of such differences were not taken into account in this study, but may be useful to investigate in future research since such differences may influence test performance.

questions were generally about common or general knowledge. Higher level tests contained progressively more questions requiring previous knowledge of a more "academic" nature.

Second, lowest level tests represented a more limited use of words and concepts than higher level tests. For example, in "matching," a word and picture usually represented identical things. Also, in "contextual paraphrase" the same word was used both in the reading selection and in the answer choice. On the other hand, in higher level tests, "semantic paraphrase" used different words to say the same or similar things. And "previous knowledge" required the use of numerous words and concepts neither presented nor necessarily implied in the reading selection.

5. The range of question tasks became broader at higher test levels. The highest concentrations of question tasks were in "probable concepts" (37%) and matching (26%) at the lowest test level. Although many questions were concentrated in "contextual paraphrase" (35%) at the intermediate test level, there were also many "probable concept" (15%), and "previous knowledge" (15%) questions. At the most advanced test level, there was an even broader distribution, i.e. 24% "semantic paraphrase," 21% "contextual paraphrase," 18% "probable concept," and 18% "previous knowledge" questions.

Consequently, at the lower grade levels, pupils could achieve adequately on subtests of reading comprehension if they could match pictures to words and could identify simple words and concepts. At the intermediate test level,pupils were being tested more on the flexibility of their vocabulary, e.g. using the same words in different contexts. At the most advanced test level pupils were tested more on the breadth of vocabulary, e.g. saying the same thing in different ways. Generally, however, it was not a question of the student having to supply the correct answer to a question. Rather, the student had to choose the right answer to the question from a number of choices which related to the question in different ways.

6. Typical distractors were similar in the lowest, intermediate and advanced test levels. At each test level the most frequent distractors were words that were grammatical answers to the question as well as somewhat related to the meaning of the correct answer, i.e. words that described a function, attribute, etc. of the correct answer. These distractors were called "associative-grammatical" and were 33% of the distractors in the lowest level tests, 30% of the distractors in the intermediate level tests and 33% of the distractors in the advanced level tests. The second most frequent type of distractors were those that fit the grammatical context of the question but were otherwise unrelated to the correct answer. Such choices were called "grammatical" and made up 20% of the distractors in tests intended for grades 4-6 and 25% of the distractors in tests intended for grades 9-14.

7. Despite the similarities among the typical distractors at the three test levels some differences did appear in the overall distribution of distractors. The percentage of "grammatical" distractors consistently increased from level to level, and so did the percentage

of "associative" distractors, e.g. "grammatical" distractors were 16% at the lowest test level, 20% at the intermediate test level and 25% at the advanced test level.

The other difference appeared when all those distractors that were words used in the reading selection were combined, no matter what other type of relationship they had with the question or correct choice, i.e. adding the number of "textual," "textual-grammatical," "textual-associative," etc. distractors for each level. Lower level tests had more distractors that were words used in the reading selection than higher level tests. The percentages were 35%, 27% and 23% in lowest, intermediate and advanced level tests respectively.

The second objective of this study was to characterize the nature of reading comprehension as tested by different test batteries.

In order to determine the tasks characteristic of each test battery, the lowest, intermediate and advanced test levels within each battery were occasionally combined. Figure 4 presents typical items for the CAT, GMRT and SAT. Tables 36-38 in Appendix A, present the percent of reading selections, questions and choices in each scale category for the CAT, GMRT and SAT. The following 6 generalizations were made on the basis of the task analysis.

1. Findings about test levels in the CAT, GMRT and SAT were similar to the findings about test levels when batteries were combined:

   a. typical reading selections were different in the lowest, intermediate and advanced level tests.

   b. the range of selection topics became broader at higher test levels.

TEST BATTERY	SELECTION	QUESTION	CHOICES
CAT	Story Science Social Studies	Contextual Paraphrase Semantic Paraphrase	Associative-Grammatical Grammatical Categorical-Grammatical
GMRT	Story Science	Previous Knowledge Probable Concept	Associative-Grammatical Grammatical
SAT	Story Science	Contextual Paraphrase Probable Concept	Associative-Grammatical

Figure 4: Typical reading comprehension items in three test batteries

c. reading selections at each higher test level included more "academic" topics.

d. typical questions were usually different in the lowest, intermediate and advanced level tests.

e. the range of question tasks became broader at higher test levels.

f. differences among distractors at the 3 test levels became clearer when test batteries were analyzed separately. Only the SAT had consistently similar distractors at the three grade levels. The GMRT and the CAT had different combinations of distractors in the 3 test levels analyzed.

For example, typical reading selections were different for the 3 test levels of the CAT, GMRT and SAT, e.g. in the lowest level CAT, "story" was the category of all the reading selections; but "story" never appeared in the intermediate and advanced CAT. In the intermediate level CAT, 67% of the selections were about science and the other 33% were about social studies. In the advanced level CAT, 40% of the reading selections were about social studies and 20% each were about social science, science and humanities. In the GMRT lowest level, 88% of the reading selections were stories and the other 12% were about science. The intermediate level GMRT had 43% "science" selections, 28% "social studies" selections and 19% "stories." At the highest level the GMRT had 33% "science" selections, 33% "humanities" selections and a few selections in a number of other subject areas.

Consequently, the CAT, GMRT and SAT had a different combination of reading selections at each level. The CAT, GMRT and SAT also differed in their combinations of reading selections, especially at the intermediate and advanced test levels.

2. Typical reading selections were somewhat different for the CAT, GMRT and SAT. For example, combining the test levels, the CAT had three frequent kinds of reading selections: 33% "story," 25% "social studies" and 25% "science."[1] Both GMRT and SAT typically had either "story" or "science" selections. The GMRT had 35% "story" selections and 31% "science" selections, while the SAT had 30% "story" and 30% "science" reading selections.

3. The CAT, GMRT and SAT differed in the number of selection categories they included. The CAT had the fewest categories, i.e. "story," "social studies," "social science," "science" and "humanities." The GMRT had six categories, i.e. "story," "language," "social studies," "social science," "science" and "humanities." The SAT had the most reading selection categories, i.e. "riddle," "story," "language," "math," "social studies," "social science," "science" and "humanities."

Despite the differences among test batteries in the topics of reading selections at the intermediate and advanced test levels, the reading selections all tended to be about school subjects. As noted earlier, there were essentially no selections that resembled other than

[1] The breadth of reading selections in the CAT may be deceiving. The CAT had the fewest selections of any battery, e.g. the CAT had 12 selections in the entire battery compared to 58 selections for the GMRT and 95 for the SAT. Hence, even a few selections in one topic became a rather high percentage.

"school type" reading material, e.g. newspaper articles, advertisements, recipes, job applications, etc. Furthermore, the selection topics resembled only a narrow range of school subjects, i.e. science and social studies. Reading selections in literature, math, art , etc. were few in number.

4. Typical questions were also somewhat different for each test battery. Characteristically, the CAT asked either "contextual paraphrase" (28%) or "semantic paraphrase" (21%) questions (see Rating Scale for Questions, p. 86, for a definition and example of question categories). The GMRT characteristically asked "previous knowledge" (22%), and "probable concept" (21%) questions. On the SAT, 34% of the questions were "contextual paraphrase" and 22% of the questions were "probable concept."

Possibly the question structure, e.g. cloze-like blanks in the reading selection, separate questions following the reading selection, influenced the question task. The CAT which always had separate questions primarily had "contextual or semantic paraphrase" tasks. However, the CAT had other tasks as well, i.e. 17% "probable concept", 12% "recognition," 7% each of "grammatical paraphrase," "definite concept," and "previous knowledge."

The GMRT, which on the intermediate and advanced test levels always had cloze-like blanks primarily had "previous knowledge" or "probable concept" tasks. However, the GMRT also required other tasks, i.e. 17% "matching," 13% "language concept," 10% contextual paraphrase," 9% "semantic paraphrase" and 7% "grammatical paraphrase."

Finally, the SAT, which had both cloze-like and separate questions seemed to require tasks most characteristic of both the CAT, e.g. "contextual paraphrase" and the GMRT, e.g. "probable concept." The SAT also had 14% "previous knowledge," 14% "semantic paraphrase," 9% grammatical paraphrase," 6% "language concept," 1% "definite concept" and less than 1% "recognition" tasks.

Thus, it appeared that the different types of questions, e.g. cloze-like blanks, separate questions, were used to create almost all of the defined tasks, e.g. "contextual paraphrase," "previous knowledge." However, certain types of tasks seemed most characteristic of certain types of questions, e.g. cloze-like blanks were characterized by requiring the use of general or academic knowledge not presented in the reading selection. Separate questions were characterized by tasks requiring the use of words stated in the reading selection in a different context, or the use of different words to restate ideas presented in the reading selection.

5. Choices were also somewhat different in the CAT, GMRT and SAT.

CAT distractors were most broadly distributed, e.g. 27% were "associative-grammatical," 24% were "grammatical" and 24% were "categorical-grammatical" (see Rating Scale for Choices, p. 89, for a definition and example of choice categories.)

GMRT distractors were generally either "associative-grammatical (30%) or "grammatical" (27%). SAT distractors were generally "associative-grammatical(34%).

Most distractors in the CAT, GMRT and SAT were grammatical answers to the questions posed. All CAT distractors were grammatical answers to the question. However 78 GMRT distractors and 28 SAT distractors were not grammatical answers to the question. When ungrammatical distractors were used to answer questions they formed odd-sounding sentences (see Appendix B). Inappropriate distractors fell into 4 categories:

a. simple grammatical error - the distractor did not agree with the number or tense of words in the question. For example: "The values of such reinforcement induces the student... (SAT, High School, Q. 19)."

b. category error - the distractor represented the wrong part of speech, e.g. the question called for a noun, but the distractor was an adjective. For example: "To receive the money, he must show proper own (GMRT, Survey D, Q. 31)."

c. feature error - the distractor represented a semantic anomaly, e.g. the question called for an animate subject, but the distractor was inanimate. For example: "Pete is a house (SAT, Primary 1, Q. 35)."

d. reality error - Awareness of "reality" made the distractor seem inappropriate. For example: "The children were very empty (GMRT, Survey D, Q. 1)."

Many of the grammatical distractors also had "association value" to the correct answer. Miller (1963) described the word-association

Note: Distractors are underlined

studies which demonstrated that consistencies existed in the types of associations different people have to given words. Studies like Woodrow and Lowell's (1916) tabulation of the relative frequencies as well as categories of word associations for children and adults suggests a possible means of investigating relative difficulty of a set of distractors. For example, distractor sets may be compared by the sum of relative frequencies of associations, or by the frequency of categories of associations, e.g. if the correct answer were "table" and the distractors were "furniture (superordinate)" "eat (verb)" and "able (assonance)" a relative difficulty score might be obtained by adding the relative frequencies from the Woodrow-Lowell list: 3.7 (table-furniture), 10.2 (table-eat) and 0.43 (table-able) = 14.33. In this manner it might be possible to systematize the combination of distractors rather than continuing the present rather random and intuitive procedure. Furthermore, if identification of differences and sources of difficulty of distractor sets becomes possible, diagnosis of pupil errors that result from particular combinations of distractors may also become possible. Such diagnosis may help teachers provide pupils with more direct instruction as well as more specific exercises.

Other distractors represented the same kind of objects, events, etc. as the correct answer. What relationship a particular type of distractor had to test levels or item difficulties was not clear from the results of analyses conducted here.

6. The choices were different in the lowest, intermediate and advanced test levels of the CAT, GMRT, and SAT. Both CAT and GMRT distractors seemed more related to the selection, question and correct choice at the lowest level than at the higher test levels. For example, many of the lowest level CAT distractors were grammatical answers to the question as well as "associative" to the correct choice (35%), or were a combination of grammatical answers to the question, the same kind of "object" as the right choice and also in the reading selection (30%). Many of the intermediate level distractors were "categorical-grammatical" (40%), and many of the advanced level distractors were simply "grammatical" (30%).

SAT distractors showed an opposite trend. Distractors in the lowest and intermediate level tests were usually "associative-grammatical." Distractors on the highest level test were either textual-categorical-grammatical" or "associative-grammatical."

Thus it appeared that while the CAT and GMRT shared a similar pattern of distractors, i.e. using more words from the reading selection at the lowest test level than at either of the higher levels, the SAT had an opposite trend, i.e. using more words from the reading selection at the highest level than either of the lower test levels.

The third objective of this study was to identify the factors that may contribute to difficulty of tested comprehension.

Correlations among empirical difficulty scores -- the criterion of difficulty in the present study -- and task ratings were not possible since the task ratings were descriptive and not quantitative. However, two major observations were made about sources of item difficulty during the rating of test items.

1. Generally it appeared from the task ratings that either selections, questions or choices may be the sources of item difficulty. Items that were passed by only a small percentage of the pupils in the try-out population contained one of the following:

a. selections that had unclear or uncommon information

b. questions that required knowledge of specific facts or ideas

c. distractors that seemed to be appropriate answers to the question.

For example, in the GMRT, Survey F, the meaning of the selection empirically found most difficult, i.e. the questions with the selection were passed by an average of about 20% of the try-out population, was unclear. The selection was rated as "humanities" by the raters more by process of elimination than by a conviction that it represented philosophy.

> The objects of science, like the direct objects of the arts, are an order of relations which serve as tools to ____50____ immediate havings and beings. Goods, objects with ____51____ of fulfillment are the natural fruition of the discovery and employment of means when the connection of ends with a sequential order is ____52____.
>
> 50. effect prevent reduce export replace
>
> 51. enjoyment thoughts uses ends qualities
>
> 52. weakened required judged determined lost

Note: The correct answer is underlined.

Furthermore, what the questions were testing was also difficult to evaluate.[1] According to the Rating Scale for Questions, questions 50-52 were rated as "probable concepts." Again raters picked this question category more by a process of elimination than by a clear understanding of what was being asked. The distractors were generally "grammatical," "associative" or "associative-grammatical" except for the distractor "ends" which was also used in the reading selection.

The SAT, Intermediate I, Question 50 demonstrates a difficulty that seemed to be related more to the choices than to the selection or question. The raters judged the selection as "science" with no difficulty.

> Cattle, sheep, goats, antelope, and deer are similar in many ways. They all have hooves and may have horns. Also, they all have a fourfold stomach. Their food is swallowed in haste and is then returned to the mouth a little at a time to be chewed methodically before it is transferred to the other sections of the stomach for gradual digestion. In this respect these ruminants, or cudchewers, are alike. One major difference is in the horns. Cattle have horns with cores composed of honeycombed bone. The horns of antelope are practically solid bone, whereas the antlers of deer are true bone. The deer shed their antlers every year in the way a deciduous tree sheds its leaves, a detail in which they are unique.
>
> 50. The best title for this paragraph would be ________________.
>
> a. The Ruminants　　　　c. Horn Structure in Animals
> b. How Many Stomachs?　d. Deer, Sheep, and Cattle

Note: The correct answer is underlined.

[1] The average difficulty score for these questions was 19.8%. When a question has 5 answer choices, each choice has a 20% probability of being picked by chance. Thus, it would appear that in an item with a selection which was meaningless, a question which was totally ambiguous, and distractors which were neutral, each choice would be picked by 20% of the testees.

The question was rated as "contextual paraphrase" since the word "ruminants" was used in the reading selection in another context. In this test item the source of difficulty seemed to be the choices especially "d", i.e. distractor "d" was rated as "textual-categorical-grammatical."[1] In a sense, distractor "d" was almost a definition or an illustration of the correct answer and could easily have been substituted for the correct answer. Distractors "b" and "c" also related to the correct answer in that they included "attributes" of ruminants which were touched upon in the reading selection. This test item was answered correctly by only 11% of the pupils in the standardization population (Kelley, et al, 1966, p. 48).

2. Generally raters seemed to have greater difficulty in identifying appropriate ratings for selections and questions of ambiguous items which were passed by a smaller percentage of pupils. For example, as noted above, in such items ratings usually were made by the process of elimination.

As illustrated in the comparison of test levels, items in higher level tests seemed to become more difficult because they were based on reading selections about more academic or obscure subjects and required previous knowledge of specialized subject matter as well as broader vocabularies.

Possibly the aspects of reading selections that bring about item difficulty, e.g. clarity, generality, abstractness, could be quantified

[1] Although the words in the choice were not exactly in the same order as in the reading selection they still all appeared close together and were thus also rated "textual."

and subsequently correlated with difficulty scores. For example, a number of raters might be asked to rate reading selections by a "semantic differential". Semantic differentials could measure ideational, language and affective characteristics. A sample of three semantic differentials is presented in Appendix C.

The fourth objective of this thesis was to characterize _the nature of tested reading comprehension._

1. Three major conclusions have already been presented about the nature of tested reading comprehension:

   a. Tests of comprehension intended for grades 1-2, 4-6 and 9-14 characteristically had different reading selections and questions. Selections at the lowest test level were usually stories about common experiences, people or events, while selections at higher test levels were usually about science, history or humanities . Questions on the lowest level tests asked general information or required the matching of words to corresponding pictures. Intermediate level questions required the use of a limited number of words in different contexts. The advanced level tests required restating ideas, using words in different contexts as well as knowing "concepts" especially in science, social studies and the humanities.

   b. The _CAT_, _GMRT_ and _SAT_ included most types of selections, questions and choices identified by the rating scales, but they differed characteristically in the distributions of selection, question and choice ratings. The CAT, GMRT and SAT generally had "story" selections at the lowest test level and science selections at the higher levels. However, the SAT had more selection types than the GMRT and CAT. The CAT and GMRT had a large percentage of selections about "humanities" at the highest level while the SAT did not.

CAT questions were more of a "paraphrase" type, i.e. using words presented in the selection in different contexts, or restating ideas presented in the reading selection. GMRT questions were more of a "conceptual" type, i.e. using either general or specific information not stated in the reading selection.

While words from the selections of the lowest level CAT and GMRT were frequently distractors, words from the selections of the intermediate and advanced CAT and GMRT were seldom distractors. On the SAT, words from the selection were more often distractors at the higher than at the lower test levels.

c. Item difficulty seemed to be related to the lack of clarity in the reading selections, the amount of uncommon or academic information required by the questions, and the similarity of meaning between the correct choice and the distractors. A rough indication of item difficulty seemed to be the difficulty raters had in categorizing test items.

These conclusions suggest that reading comprehension test items especially at higher test levels could be testing "information" and "skills" that related to other school subjects as well. In characterizing reading comprehension, it seemed appropriate to establish the unique qualities of reading comprehension test items. Toward this end reading comprehension test items and items of other disciplines, e.g. science, social studies, were compared. On the basis of this comparison another conclusion was reached.

2. Reading comprehension test items closely resembled test items for other school subjects such as science and social studies.

To illustrate the similarity between comprehension test items and test items from other school subjects a total of 8 test items were selected from the social studies, science, word meaning, paragraph meaning, i.e. comprehension, and mathematics subtests of Stanford Achievement Tests.[1]

The reader is requested to read each of the following test items carefully, to establish the kind of "information" or "skill" needed to answer the questions and, consequently, to determine which school subject, i.e. social studies, science, word meaning, paragraph meaning, or mathematics the following items test:

[1] The paragraph meaning subtest of the Stanford Achievement Test was the reading comprehension test chosen for this exercise for the following reasons:

a. the paragraph meaning subtest (SAT) tended to contain qualities of both the CAT and GMRT (see preceding readability and task analyses of comprehension subtests).
b. publishers of the Stanford Achievement Tests generously provided the subtests for science, social studies, word knowledge, etc.
c. intercorrelations of subtest scores were readily available in test manuals.

1. "O beautiful for heroes proved, In liberating strife..." These heroes were probably the heroes of __________.

   a. 1914
   b. 1861
   c. 1776
   d. 1898

2. From 1850 to 1880, Virginia City held a prominent place in the history of silver and gold mining. Its fabulous production of silver and gold has left a tremendous impression on all who ever heard of it. This production played an important role in financing the Union during the ________________.

   a. War between the States
   b. Revolutionary War
   c. War of 1812
   d. Mexican War

3. Costa Rica is south of the United States. Since Costa Rica is in Central America, the United States is ________ of Central America.

   a. north
   b. south
   c. part
   d. in the middle

4. A boy has to walk directly west in going from his home to school. To come home quickly, he should walk______.

   a. north
   b. west
   c. south
   d. east

5. Ruth wasn't upset by the little old man. Although he was strange, she was rather pleased by him. She thought he was __________.

   a. wicked
   b. fearful
   c. quaint
   d. dirty

6. A person who attempts to change or improve conditions is called a _______________.

   a. coward
   b. conservationist
   c. reformer
   d. conservative

7. A country is measured and mapped by means of trigonometry-the branch of mathematics dealing with the measurement of triangles. When we know the length of one side of a triangle and the size of the two angles at its ends, we have the information that will give us the length of the other ____I_______ of the triangle and the size of the third ______II_____of the triangle.

   I a. side
     b. three sides
     c. four sides
     d. two sides

   II a. arc
     b. altitude
     c. base
     d. angle

8. Suppose that we knew the formula for the area of a triangle. We could use it to find formulas for the area of ________________.

   a. rectangles, squares, and paralellograms, but not trapezoids
   b. rectangles, squares, parallelograms, and trapezoids
   c. rectangles and squares, but not parallelograms or trapezoids
   d. none of the above

Answers:

1. Social Studies subtest, item 29
   Stanford Achievement Test, Intermediate 2, Form X

2. Paragraph Meaning subtest, item 2
   Stanford Achievement Test, Intermediate 2, Form Y

3. Paragraph Meaning subtest, item 12
   Stanford Achievement Test, Intermediate 1, Form X

4. Science subtest, item 20
   Stanford Achievement Test, Intermediate 1, Form X

5. Paragraph Meaning subtest, item 17
   Stanford Achievement Test, Intermediate 2, Form X

6. Word Meaning subtest, item 41
   Stanford Achievement Test, Intermediate 2, Form X

7. Paragraph Meaning subtest, items 47 and 48
   Stanford Achievement Test, Intermediate 1, Form X

8. Mathematics subtest, item 49
   Stanford Achievement Test, High School, Form X

The selections, questions and choices of "paragraph meaning" -- reading comprehension -- test items were very similar to the selections, questions and choices of test items from other school subjects such as social studies, science, word meaning and mathematics.

The investigation of similarity among subtests was pursued by a study of the relationship between reading comprehension test scores and the scores of tests in the other disciplines.

3. Comprehension tests seemed to be measuring the same kind of "abilities" as tests of other school subjects especially word meaning, science and social studies.

Table 9 presents correlation coefficients of Stanford Achievement Test paragraph meaning scores and scores of word knowledge, spelling, arithmetic, social studies, and science subtests. Correlation coefficients of paragraph meaning test scores and Otis I.Q. scores are also presented in Table 9.

The paragraph meaning scores correlated very highly with word knowledge (.72 to .83), science(.72 to .82) and social studies (.75 to .81) According to Commins and Fagin (1954, p. 327-328) "When a number of tests have high intercorrelations, we may assume that they are measuring to a large extent the same kinds of abilities..."

Table 9

Correlation Coefficients of Stanford Achievement Test Paragraph Meaning Scores and Scores of Word Knowledge, Spelling, Arithmetic, Social Studies, Science and Otis I.Q.

Level and Grade	Stanford Achievement Test[a]					Otis[b]
	Word Knowledge	Spelling	Arithmetic Concepts	Social Studies	Science	I.Q.
Lowest (Grade 1)	.72	.71	.60	--	--	.39
Intermediate I (Grade 4)	.82	.74	.67	.77	.82	.73
Intermediate II (Grade 5-6)	.83,.83	.70,.70	.63,.72	.79,.81	.79,.80	.70,.75
Advanced[c] (Grades 9,10, 11,12)	-- -- -- --	.65,.63, .61,.60	.65,.67, .66,.64	.79,.75, .80,.80	.75,.79, .74,.72	.81,.80, .82,.82

Note: A hyphen means that no correlations were available.

[a] Kelley, et al (1966, p. 18). The correlations were from randomly selected samples of approximately 1000 pupils per grade. The critical correlation for 1000 degrees of freedom at the .01 level of significance is .08.

[b] Kelley, et al (1966, p. 24). The correlations were from randomly selected samples of approximately 1000 pupils per grade.

[c] Gardner, et al (1965b, p. 20-21). The population from which these correlations were obtained was not specified. However, it seemed likely that the standardization population was used, i.e. 5700 pupils per grade (Kelley, et al, 1965b, p. 11).

The consistently high correlations of comprehension and word knowledge test scores seemed to correspond with the earlier finding that many comprehension questions required breadth and depth of vocabulary, e.g. "matching," "contextual paraphrase," "semantic paraphrase."[1]

P. E. Vernon's (1962, p. 269) observation that all subject matter tended "to take the form of complex reading comprehension tests" seemed to apply in the reverse as well. The considerable percentage of "previous knowledge" questions on tests of comprehension suggested that pupil performance on tests of comprehension depended, in part, on the pupil's knowledge of information not stated in the reading selection. The numerous reading selections about science and social studies in tests of comprehension suggested that knowledge of science and social studies was required. The generally higher correlations of reading comprehension with social studies and science than with spelling and arithmetic seemed to corroborate this conclusion.

Although the correlations for paragraph meaning test scores with scores of spelling and arithmetic tests were somewhat lower (.60 to .74), they still showed a considerably close relationship between the tests. I.Q. scores had a relatively low correlation (.39) with paragraph meaning scores at the lowest test level. However, the correlation of I.Q. and paragraph meaning scores increased through the test levels and was .82 at the advanced test level.

[1] Breadth refers to knowing the meaning of many different words and depth refers to knowing the many meanings of a given word.

The Otis I.Q. test intended for lower elementary school grades consisted entirely of picture items and oral instructions (Otis, 1954, p. 1).[1] The reading comprehension tests analyzed at corresponding levels, i.e. grades 1-2, required reading of words, sentences and paragraphs. Thus, the two types of tests did not appear to be testing similar "abilities" and the relatively low correlations were to be expected. However, at higher grade levels sections of many "paper-and-pencil" I.Q. tests were essentially identical to reading comprehension tests. Higher level I.Q. tests generally contained some reading selections, questions about the selections, and multiple-choice answers. Thus, the two types of tests appeared to be testing some identical "abilities" and therefore, the higher correlations were to be expected.

In addition, the high correlation at higher grade levels between scores on reading comprehension and I.Q. tests may also have resulted from the interdependent validity of these tests. For example, some reading test-authors assumed that "circumstances that contribute to high or low I.Q. scores in a school population are also the main factors contributing to high or low reading scores (Gates and MacGinitie, 1970, p. 1). Thus these test authors used I.Q. scores as an "external validity criterion." Conversely, "many intelligence tests are validated against measures of academic achievement...(Anastasi, 1961, p. 190)," i.e. standardized achievement tests. The difference in correlations of

[1] The California Test of Mental Maturity intended for lower elementary grades also consisted entirely of picture items and oral instructions (Sullivan, 1963, p. 6).

I.Q. and reading test scores at higher and lower levels may also be attributed to Chall's (1967, p. 138-139) suggestion that intelligence would be more of a factor in limiting performance on advanced "aspects of reading comprehension, such as 'reading to predict outcomes,' 'making inferences,'...and the like," than on less advanced aspects such as "reading for details and following directions."

To view the relationship of reading comprehension test scores and scores of tests in other school subjects in proper perspective, a study of the relationship of scores from different reading comprehension tests was undertaken.

4. Scores of different comprehension tests did not seem to be more highly related to each other than to scores of tests in other school subjects.

Table 10 presents correlation coefficients for scores of the California Achievement Test comprehension subtest with scores of the a) California Achievement Test vocabulary subtest, spelling subtest, and and arithmetical reasoning subtest, b) California Test of Mental Maturity language and non-language I.Q.s , c) Metropolitan Achievement Test reading, i.e. comprehension, subtest, d) Iowa Tests of Basic Skills comprehension subtest and vocabulary subtest, and e) Stanford Achievement Test paragraph meaning subtest.

Correlations of California Achievement Test comprehension scores with test scores of other school subjects generally corresponded to those on the Stanford Achievement Test presented in Table 9:

Table 10

Correlations of the California Achievement Test, Comprehension Subtest with Other California Subtests, with I.Q. and with Comprehension Subtests of Other Batteries

Test	Level and Grade						
	Lowest	Intermediate			Advanced		
	1	4	5	6	10	11	12
California Achievement Test[a]							
Vocabulary	.75	.58	.79	.76	.85	.80	.85
Spelling	.67	.50	.70	.55	.71	.67	.66
Arithmetical Reasoning	.56	.60	.75	.78	.77	.78	.77
California Test of Mental Maturity[b]							
Language I.Q.	.14	.71			.77		
non-Language I.Q.	.13	.52			.63		
Metropolitan Achievement Test							
Reading (Comprehension)	--	.78[c], .85[d]			--		
Iowa Test of Basic Skills							
Comprehension	--	.87[c], .80[e]			--		
Vocabulary	--		.76[e]		--		
Stanford Achievement Test[d]							
Paragraph Meaning (Comprehension)	.62	.85			--		

Note 1: A hyphen means that no correlations were available

Note 2: The critical correlation coefficient for 70 degrees of freedom at the .01 level of significance is .30.
The critical correlation coefficient for 200 degrees of freedom at the .01 level of significance is .18.

[a] California Test Bureau (1967, pp. 43-45). The number of pupils differed for each grade and ranged from 77 to 125.

[b] California Test Bureau (1957, pp. 26-27). The number of pupils differed for each grade and ranged from 83 to 108.

[c] Finley (1963, p. 82). The number of pupils was 159.

[d] Tiegs and Clark (1963a, p. 12; 1963b, p. 13). The number of pupils ranged from 86 to 126.

[e] Garlock and Harsh (1960, p. 152). The number of pupils was 241.

a. correlations were generally high

b. the test most highly correlated with reading comprehension seemed to be word knowledge

c. I. Q. scores had a low correlation with reading comprehension scores at the lowest test level, but a relatively high corrclation at higher test levels.

Table 10 indicates that the correlation at the lowest test level of scores on California Achievement Test comprehension and on Stanford Achievement Test paragraph meaning was .62. This correlation was lower than the correlation of the California Achievement Test comprehension subtest scores to both California Achievement Test vocabulary (.75) and spelling (.67) subtest scores at that level.

At the intermediate test level the correlations among different reading comprehension subtests ranged from .78 to .83, while the correlations of reading comprehension subtests to subtests of other school subjects ranged from .50 to .79.

The study of correlation coefficients did not indicate the existence of major differences among reading comprehension test scores and test scores of word knowledge, science, social studies, and intelligence. A comparative analysis of items from various tests clarified this phenomenon. All these tests appeared to require knowledge of word meanings and uses, knowledge of general information, and knowledge of information related to selected school subjects, e.g.

social studies, science.[1] Consequently, scores of reading comprehension tests generally did not appear to tell the teacher more about pupils' "reading ability" than did scores of tests on intelligence, or other selected school subjects.

[1] There are numerous other influences on test performance which do not relate to item content and are therefore outside the topic of this thesis. For example, test characteristics such as test instructions and the conditions under which the test is administered influence performance (Klein, 1971, p. 3-4). Many pupil characteristics also influence test performance such as motivation and test-taking skill (Anastasi, 1961, 61-66; Cronbach, 1954, 181-187).

CHAPTER VI

New Tests of Reading Comprehension

Different tests of reading comprehension emphasize different stylistic elements as demonstrated by the readability analysis (Chapter IV), and different tasks, as demonstrated by the task analysis (Chapter V). Yet they all correlate highly with each other and with tests in other subject areas. Most of these tests appear to be measuring vocabulary, general intelligence, "reading" and previous knowledge of school subjects to a lesser or greater degree.

Further study of the relationship between readability and tasks in reading comprehension tests would undoubtedly be enlighting.[1] However, the information accumulated by the present analysis is sufficient to suggest some requirements of new tests of reading comprehension. The new tests would not only establish the rank of a pupil in relation to pupils in the standardization or norming population of the tests, but would provide teachers with more specific diagnostic information. Such information could be used to establish a pupil's performance level in relation to the "criterion" of expected performance and consequently also point out specific weaknesses. The new tests would include 4 major features:

[1] To establish statistically whether differences exist in the empirical difficulty of the numerous combinations of selections, questions and choices an analysis of variance approach seems most appropriate. To establish the relationship among the numerous combinations of selections, questions and choices while controlling for the number or ratio of difficult words, a covariance approach seems appropriate. Both these approaches may be combined into one analysis of covariance using empirical item difficulty scores as data, and using the combined number or ratio of difficult words in the selection, question and choices as the covariate.

1. A definition of minimum length, sentence length and hard word ratio for reading selections, questions and choices at the numerous grade or test levels.

Reading comprehension relates to long and short reading matter as well as to reading matter with many or few hard words.[1] The preceding analyses revealed that pupils at lower grade levels generally were tested by shorter reading selections with fewer hard words than pupils at higher grade levels. Yet, the most appropriate length or hard word ratio for reading matter at a given test level was not apparent. Establishing minimum "criteria" in this respect, for the grade or test levels would improve the understanding both of what reading

[1] The relationship of sentence length and "sentence complexity" has already been noted. In attempting to establish a minimum "criterion" for sentence length or "complexity" analyses such as those by Carol Chomsky (1969) of the age level at which pupils acquire understanding of certain syntactic structures may prove most useful.

Furthermore, lists of "easy" words which would be understood by selected age or grade groups are available. For example, Stone's Revision of Dale's List of 769 Easy Words includes words which most 1st graders are expected to know. The Dale List of 3000 Familiar Words includes words which most 4th graders are expected to know. Consolidation and expansion of these and similar lists could help establish a minimum "criteria" for a given grade or test level. However, in determining minimum "vocabulary" particular care should be taken not to discriminate against the segments of the population who may have a considerable "non-academic" vocabulary, but may have a limited "academic" vocabulary.

Edgar Dale and Gerhard Eichholz have been working on comprehensive lists for selected grades. Their final results have not been published however an interim report, Children's Knowledge of Words, Bureau of Educational Research and Service, The Ohio State University (1954 to 1960), was printed.

comprehension at a given level entails and of what difficulties given pupils have in reading comprehension.[1]

2. A definition of the subject of selections in reading comprehension tests.

Reading comprehension is related to all school subjects and to reading material not necessarily read in schools. But, the vocabulary and language structures used in "school" and "non-school" reading matter are not necessarily identical. Understanding reading selections about social studies for instance, does not necessarily indicate understanding of trade manuals, or contemporary literature.

If the objective of the tests is to establish how well pupils read "academic" subjects, then the tests selections about social studies, science, and humanities for example, are most appropriate. However, if the objective of the test is to establish how well pupils understand "non-academic" reading, excerpts from newspapers, magazines, etc. would seem more appropriate. And if the objective of the test is to establish how well pupils cope with vague or meaningless reading, such reading selections would be appropriate.

[1] Glaser and Cox (1968, p. 545) in contrasting currently used achievement tests with "criterion-referenced" tests explained that the currently used tests "need provide little or no information about the degree of proficiency exhibited by the tested behaviors in terms of what the individual can do. They tell that one pupil is more or less proficient than another, but do not tell how proficient either of them is with respect to the subject-matter tasks involved." On the other hand, criterion-referenced tests assess "The degree to which an individual's achievement resembles desired performance at any specified level along the continuum of attainment...."

It would be useful to determine the grade level at which particular topics could most appropriately be introduced or dropped in sequential testing.[1] For example, at the lowest grade level selections are mainly "stories." It is unclear whether other topics such as social studies or science might not also be introduced at the lowest grade level.[2] As revealed in the preceding analysis, the percentage of stories about common events, people or experiences at the highest test level is low. Yet, "stories" are a popular and frequent form of adult reading both in and out of school, and therefore, may appropriately be included in advanced level tests.

3. A definition of the tasks necessary for supplying correct answers to questions.

The preceding task analysis has identified types of questions found on current comprehension tests. Generally, reading comprehension questions require either "paraphrase" or "concept" tasks.

[1] Generally, reading matter in the 1st and 2nd grade is concentrated in school readers which contain mostly "stories." However, pupils in the 1st and 2nd grade are also taught some social studies and science. They may even do some reading in school about more "academic" topics. This leads to the question of curricular validity of tests which is the correspondence between test and curriculum content (Kelley, *et al*, 1966, p. 23). Usually it is expected that the test is designed according to the curriculum. However, Klein, (1970, p.2) suggested that it is not uncommon for educators to modify a curriculum to correspond with tests. Thus, it seems appropriate for test authors as well as educators to study these questions.

[2] The lowest level *SAT* had approximately 4% "science" selections.

"Paraphrase tasks" require pupils to pick answers which are restatements of information explicitly given in the reading selections. "Restatement" is possible in a number of ways. For example, sometimes the answer is a picture representing the word(s). Sometimes the word is grammatically changed, e.g. different tense or number. Other times, different words with the same meaning are used. An additional influence on "paraphrase tasks" is the context in which the information is presented. Sometimes the context of information given in the reading selection is essentially the same as the context of the same information in the answer, but not always.

Figure 5 presents the "paraphrase tasks" found on the analyzed reading comprehension tests. To summarize briefly, the following 6 "paraphrase tasks" were identified:

a/b. matching/selecting - the information was stated in a word(s); but the answer was a picture representing the same thing.[1]

c. recognizing - the same word(s) was used in the reading selection and answer. The contexts of the selection and answer were also essentially the same.

d. contextual paraphrase - the same word(s) was used in the reading selection and answer. However, the context of the

[1] Due to the small number of picture answers, all questions that required picture-word matching were put into one category. However, there were really two types of items. In one type, matching, pictures represented the words exactly. In the other type, selecting, the picture either added to or omitted from information described by the words.

FORM OF INFORMATION	CONTEXT OF INFORMATION	
	Identical	Different
Picture-Word	matching	selecting
Identical Word	recognizing	contéxtual paraphrase
Grammatically Different	grammatical change[a]	grammatical paraphrase
Semantically Different	semantic change[a]	semantic paraphrase

Figure 5: Paraphrase tasks in reading comprehension test items

[a] These tasks did not appear on' the reading comprehension tests analyzed.

information in the answer was different from the context of the same information in the reading selection.

e. grammatical paraphrase - the word(s) used in the selection was grammatically different, e.g. tense, number, from the "same" word(s) in the answer. The contexts were also different.

f. semantic paraphrase - the word(s) used in the selection were different from the words used in the answer; but, they both meant the same thing. The contexts were also different.

Two types of questions do not appear in the reading comprehension tests analyzed:[1]

a. grammatical change - the word(s) used in the selection is grammatically different from the "same" word(s) in the answer. The context is essentially the same.

b. semantic change - the word(s) used in the selection is different than the word(s) used in the answer; but they both mean the same thing. The context is essentially the same.

[1] The value of such items lies in the possibility that they may facilitate the transition of learning to cope with progressively harder reading comprehension questions. For example, it may be that if matching is the simplest question task, selecting may be a bit more difficult, then recognizing, contextual paraphrase, grammatical change, grammatical paraphrase, and so forth would become progressively more difficult.

"Concept tasks" require pupils to choose answers which represent general or academic knowledge. The concepts are never explicitly stated in the reading selection. However, the selection gives some hints or cues. For example, sometimes generally known concepts are cued by descriptions of their features. Other times generally known concepts are cued by syntactic implications, e.g. colloquialisms, idioms. On numerous occasions "academic" concepts are cued by their features, or by related concepts. An additional influence on concept tasks is the probability or certainty with which an answer is identified. For example, sometimes only one answer fits the cues. Other times one answer fits the cues only a little bit better than another.

Figure 6, presents the 4 "concept tasks" found on the three reading comprehension tests analyzed:

a. definite concept - features of the concept which are given in the reading selection clearly identify only one answer.

b. probable concept - features of the concept which are given in the reading selection imply that one answer is probably better than another.

c. (probable) language concept - the language structures in the reading selection suggest that one answer is probably better than another. This category generally applies only to questions in the form of cloze-like blanks.

d. (definite) previous knowledge - previous knowledge of "academic" facts clearly identifies one and only one answer.

	CUES TO THE CORRECT ANSWER		
PROBABILITY OF THE CHOICE	Stated Features	Syntactic Implications	Previous Knowledge
Definite	definite concept	(definite) language concept[a]	(definite) previous knowledge concept
Probable	probable concept	(probable) language concept	(probable) previous knowledge concept[a]

Figure 6: Concept tasks in reading comprehension test items

[a] These tasks did not appear on the reading comprehension tests analyzed.

Two types of questions do not appear in the reading comprehension tests analyzed:[1]

a. (definite) language concept - the language structures in the reading selection definitely imply only one answer.

b. (probable) previous knowledge - previous knowledge of "academic" facts suggests one answer more than another but neither definitely.

Generally, if the objective of the test is to establish how well pupils manipulate explicitly stated information, then "paraphrase" questions are appropriate. However, if the objective of the test is to establish how well pupils manipulate "general concepts," then "definite concept" or "probable concept" questions are more appropriate. If the objective is to establish pupils' fluency in English, "language concept" questions are more appropriate. And finally, if the objective is to establish pupils' knowledge of academic facts, then "previous knowledge" questions seem more appropriate. Whether questions testing language fluency or previous knowledge belong on tests of reading comprehension is not clear. Apparently achievement tests in English test language fluency, and achievement tests in specific school subjects test knowledge of facts. The inclusion of such items on tests of reading comprehension has received the following criticism from Marks and

[1] Again, the value of such items would lie in the possibility that they could facilitate the transition of learning to cope with progressively harder reading comprehension questions.

Noll (1967, p. 346):

> Our intuitive notion of the comprehension task leads us to conclude that tests where scores are unduly influenced by specific previous knowledge or response biases are invalid measures of this ability.

Similarly Guttman (1965) differentiated between "achievement" type items that would require previous knowledge of facts and "analytic ability" type items which would require the ability to analyze or manipulate given information.

Finally, determining whether or not a given sequence of questions through the many test levels facilitates better performance may prove useful for both testing and teaching.

4. A definition of the character of distractors in tests of reading comprehension.

Distractors were initially introduced into the testing of reading comprehension essentially to facilitate scoring and not to influence item difficulty. However, they generally do affect item difficulty and therefore, may obscure rather than clarify the meaning of reading comprehension test scores. Twelve types of distractors were identified (see Rating Scale for Choices, p. 89). Distractor combinations were often established during test construction by giving the questions to a trial population in open-ended form. The most frequent errors made by the trial population were later made distractors in the multiple-choice form of that test (California Test Bureau, 1957, p.6). However, the nature of the most frequent errors was not analyzed and their effect on item difficulty remained unknown. But, on the basis of the distractor types identified in the preceding task analysis, it should be possible to diagnose the types of errors pupils make consistently and to control distractor difficulty.

In conclusion, if 'reading comprehension" is to be a meaningful construct in teaching and testing, it seems to require a clear definition. Otherwise instruction of "reading comprehension" is simply a replication of instruction in science, history, or vocabulary. And, testing "reading comprehension" is simply a combination of testing intelligence and numerous school subjects. Each test should focus on a specific objective and reduce the influence of extraneous factors. For example, tests in science could be simply worded reducing the influence of word knowledge. Tests in reading comprehension could provide all the subject matter information needed, reducing the influence of previous knowledge. Furthermore, if test-authors identify the particular combination of "selections," "questions" and "choices" which they consider "comprehension," the construct may develop defined features. For example, one test-author might focus on "story"selections, "paraphrase" tasks and "grammatical" distractors. Another test-author might prefer "academic" selections, "concept" tasks and "textual" distractors, and so on. Specifying objectives in this manner may help test-authors in constructing their tests. Descriptions of items may also permit teachers and administrators to decide more quickly and more knowledgeably if given tests are valid instruments for their purposes.

Test-authors could also greatly facilitate the diagnosis and possibly treatment of pupils who fail tests by specifying how item difficulty is increased. For example, one test-author may increase the proportion of difficult words. Another test-author may increase the ambiguity of the question and so on.

Finally if literacy is a national priority and the attempt to teach almost all citizens to read is continued, the "normal distribution" model used in the design of current reading comprehension tests is inappropriate. According to this model prearranged proportions of the population are designated as doing very well, sufficiently well and "failing" on the test. Thus, a sizeable proportion of the national population achieves below "grade level" by definition.

However, the use of the "criterion" model suggested above would not condemn a considerable portion of the population to failure. By defining the "criteria" of reading comprehension, this model would facilitate not only a more meaningful evaluation of reading comprehension but would also facilitate the teaching of reading.

APPENDIX A

SUPPLEMENTARY DATA TABLES

LIST OF TABLES IN APPENDIX A

Table 1

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Choices Accompanying a Question by Test

Test	Number of Questions	Mean number of Choices	Standard Deviation	Minimum number of Choices	Maximum number of Choices	Range
<u>CAT</u>						
1. Lower Primary	15	2.53	0.99	1.00[a]	4.00	3.00
2. Elementary	30	4.00	0.0	4.00	4.00	0.0
3. Advanced	45	4.00	0.0	4.00	4.00	0.0
<u>GMRT</u>						
4. Level A	34	4.00	0.0	4.00	4.00	0.0
5. Level D	52	5.00	0.0	5.00	5.00	0.0
6. Level F	52	5.00	0.0	5.00	5.00	0.0
<u>SAT</u>						
7. Primary I	38	4.00	0.0	4.00	4.00	0.0
8. Intermediate I	60	4.00	0.0	4.00	4.00	0.0
9. Intermediate II	64	4.00	0.0	4.00	4.00	0.0
10. High School	65	4.00	0.0	4.00	4.00	0.0

[a] One item was open-ended, e.g. "Write a word that begins with d (CAT, Lower Primary, Q. 4)." Since any word beginning with "d" would have been acceptable and since a word beginning with "d" was presented in the choices of the immediately preceding item, this item was treated as if it had one choice. Three other items were treated as if they had one choice. In each case, the pupils were required to write the missing letters in a mutilated word when the correct form of that word was provided as a standard of comparison, e.g. CAT, Lower Primary, Q.1,2,5.

Table 2

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Words in the Reading Selections by Test

Test	Number of Selections	Mean number of Words	Standard Deviation	Minimum number of Words	Maximum number of Words	Range
CAT						
1. Lower Primary	4	23.00	5.83	17.00	28.00	11.00
2. Elementary	3	198.00	60.22	156.00	267.00	111.00
3. Advanced	5	418.60	161.83	229.00	637.00	408.00
Total	12	231.58	205.47	17.00	637.00	620.00
GMRT						
4. Level A	16	20.38	8.64	7.00	35.00	28.00
5. Level D	21	42.81	9.77	28.00	71.00	43.00
6. Level F	21	57.33	20.16	34.00	109.00	75.00
Total	58	41.88	20.34	7.00	109.00	102.00
SAT						
7. Primary I	33	17.06	7.00	8.00	42.00	34.00
8. Intermediate I	24	64.79	38.30	13.00	161.00	148.00
9. Intermediate II	25	67.04	31.29	12.00	127.00	115.00
10. High School	13	137.38	102.34	39.00	383.00	344.00
Total	95	58.74	58.78	8.00	383.00	375.00

Table 3

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Sentences in the Reading Selections by Test

Test	Number of Selections	Mean number of Sentences	Standard Deviation	Minimum number of Sentences	Maximum number of Sentences	Range
CAT						
1. Lower Primary	4	3.75	1.26	2.00	5.00	3.00
2. Elementary	3	13.00	2.65	11.00	16.00	5.00
3. Advanced	5	20.60	10.50	12.00	36.00	24.00
Total	12	13.08	9.96	2.00	36.00	34.00
GMRT						
4. Level A	16	2.38	1.09	1.00	4.00	3.00
5. Level D	21	2.90	0.70	2.00	4.00	2.00
6. Level F	21	2.62	0.92	1.00	4.00	3.00
Total	58	2.66	0.91	1.00	4.00	3.00
SAT						
7. Primary I	33	3.24	0.94	2.00	6.00	4.00
8. Intermediate I	24	3.92	2.14	2.00	10.00	8.00
9. Intermediate II	25	3.68	1.70	1.00	9.00	8.00
10. High School	13	5.31	2.98	2.00	12.00	10.00
Total	95	3.81	1.93	1.00	12.00	11.00

Table 4

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Spache Words in the Reading Selections by Test

Test	Number of Selections	Mean number of Non-Spache Words	Standard Deviation	Minimum number of Non-Spache Words	Maximum number of Non-Spache Words	Range
CAT						
1. Lower Primary	4	1.25	0.96	0.0	2.00	2.00
2. Elementary	3	55.67	18.23	36.00	72.00	36.00
3. Advanced	5	124.20	37.61	81.00	182.00	101.00
Total	12	66.08	60.57	0.0	182.00	182.00
GMRT						
4. Level A	16	2.19	1.64	0.0	5.00	5.00
5. Level D	21	11.52	4.77	4.00	21.00	17.00
6. Level F	21	20.14	6.19	11.00	33.00	22.00
Total	58	12.07	8.58	0.0	33.00	33.00
SAT						
7. Primary I	33	0.70	0.95	0.0	3.00	3.00
8. Intermediate I	24	14.04	10.25	2.00	48.00	46.00
9. Intermediate II	25	17.44	10.05	6.00	41.00	35.00
10. High School	13	42.92	30.90	8.00	100.00	92.00
Total	95	14.25	18.83	0.00	100.00	100.00

Table 5

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Dale-Chall Words in the Reading Selections by Test

Test	Number of Selections	Mean number of Non-D-C Words	Standard Deviation	Minimum number of Non-D-C Words	Maximum number of Non-D-C Words	Range
CAT						
1. Lower Primary	4	0.0	0.0	0.0	0.0	0.0
2. Elementary	3	29.67	15.01	15.00	45.00	30.00
3. Advanced	5	108.60	33.08	70.00	154.00	84.00
Total	12	52.67	54.90	0.00	154.00	154.00
GMRT						
4. Level A	16	0.38	1.02	0.0	4.00	4.00
5. Level D	21	5.67	4.92	0.0	21.00	21.00
6. Level F	21	15.90	5.58	7.00	27.00	20.00
Total	58	7.91	7.81	0.0	27.00	27.00
SAT						
7. Primary I	33	0.03	0.17	0.0	1.00	1.00
8. Intermediate I	24	8.62	8.07	0.0	26.00	26.00
9. Intermediate II	25	11.36	6.68	2.00	22.00	20.00
10. High School	13	33.69	27.53	4.00	106.00	102.00
Total	95	9.79	15.41	0.0	106.00	106.00

Table 6

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Words in the Questions by Test

Test	Number of Questions	Mean number of Words	Standard Deviation	Minimum number of Words	Maximum number of Words	Range
CAT						
1. Lower Primary	15	12.53	3.94	7.00	22.00	15.00
2. Elementary	30	18.23	14.43	4.00	42.00	38.00
3. Advanced	45	13.64	5.70	4.00	33.00	29.00
Total	90	14.99	9.58	4.00	42.00	38.00
GMRT						
4. Level A	34	4.94	3.81	0.0[a]	13.00	13.00
5. Level D	52	16.38	5.91	3.00	29.00	26.00
6. Level F	52	25.02	10.32	7.00	47.00	40.00
Total	138	16.82	10.80	0.0	47.00	47.00
SAT						
7. Primary I	38	4.90	1.41	3.00	9.00	6.00
8. Intermediate I	60	18.85	8.81	4.00	49.00	45.00
9. Intermediate II	64	18.81	9.94	3.00	49.00	46.00
10. High School	65	17.52	8.28	5.00	37.00	32.00
Total	227	16.12	9.66	3.00	49.00	46.00

[a] Some items had no questions. There was a selection in the form of a description or direction which indicated what the pupil was expected to do (see footnote "c" on Table 6 in Text, p. 57).

Table 7

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Spache Words in the Questions by Test

Test	Number of Questions	Mean number of Non-Spache Words	Standard Deviation	Minimum number of Non-Spache Words	Maximum number of Non-Spache Words	Range
CAT						
1. Lower Primary	15	0.47	0.52	0.0	1.00	1.00
2. Elementary	30	4.17	3.20	0.0	10.00	10.00
3. Advanced	45	4.44	1.89	1.00	8.00	7.00
Total	90	3.69	2.69	0.0	10.00	10.00
GMRT						
4. Level A	34	0.35	0.69	0.0	3.00	3.00
5. Level D	52	4.64	2.94	0.0	13.00	13.00
6. Level F	52	8.77	4.15	3.00	20.00	17.00
Total	138	5.14	4.53	0.0	20.00	20.00
SAT						
7. Primary I	38	0.16	0.44	0.0	2.00	2.00
8. Intermediate I	60	4.17	2.15	0.0	12.00	12.00
9. Intermediate II	64	5.03	2.65	0.0	12.00	12.00
10. High School	65	6.05	3.10	1.00	14.00	13.00
Total	227	4.28	3.13	0.0	14.00	14.00

Table 8

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Dale-Chall Words in the Questions by Test

Test	Number of Questions	Mean number of Non-D-C Words	Standard Deviation	Minimum number of Non-D-C Words	Maximum number of Non-D-C Words	Range
CAT						
1. Lower Primary	15	0.07	0.26	0.0	1.00	1.00
2. Elementary	30	3.43	3.34	0.0	10.00	10.00
3. Advanced	45	3.73	1.80	0.0	10.00	10.00
Total	90	3.02	2.65	0.0	10.00	10.00
GMRT						
4. Level A	34	0.0	0.0	0.0	0.0	0.0
5. Level D	52	2.23	2.39	0.0	10.00	10.00
6. Level F	52	6.58	2.68	0.0	11.00	11.00
Total	138	3.32	3.47	0.0	11.00	11.00
SAT						
7. Primary I	38	0.0	0.0	0.0	0.0	0.0
8. Intermediate I	60	2.28	2.23	0.0	9.00	9.00
9. Intermediate II	64	3.28	2.26	0.0	9.00	9.00
10. High School	65	4.78	2.92	1.00	13.00	12.00
Total	227	2.90	2.78	0.0	13.00	13.00

Table 9

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Words in the Choices by Test

Test	Number of Choices	Mean number of Words	Standard Deviation	Minimum number of Words	Maximum number of Words	Range
CAT						
1. Lower Primary	38	1.68	1.51	1.00	7.00	6.00
2. Elementary	120	2.10	1.33	1.00	6.00	5.00
3. Advanced	180	3.55	2.93	1.00	19.00	18.00
Total	338	2.82	2.46	1.00	19.00	18.00
GMRT						
4. Level A	136	0.35	0.84	0.0[a]	3.00	3.00
5. Level D	260	1.00	0.06	1.00	2.00	1.00
6. Level F	260	1.00	0.0	1.00	1.00	0.0
Total	656	0.87	0.46	0.0	3.00	3.00
SAT						
7. Primary I	152	1.01	0.11	1.00	2.00	1.00
8. Intermediate I	240	1.31	0.71	1.00	5.00	4.00
9. Intermediate II	256	1.36	0.76	1.00	5.00	4.00
10. High School	260	2.16	2.43	1.00	14.00	13.00
Total	908	1.52	1.47	1.00	14.00	13.00

[a] Many items in this test had picture choices with no words in them.

Table 10

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Spache Words in the Choices by Test

Test	Number of Choices	Mean number of Non-Spache Words	Standard Deviation	Minimum number of Non-Spache Words	Maximum number of Non-Spache Words	Range
CAT						
1. Lower Primary	38	0.08	0.27	0.0	1.00	1.00
2. Elementary	120	0.75	0.60	0.0	2.00	2.00
3. Advanced	180	1.42	0.96	0.0	6.00	6.00
Total	338	1.03	0.91	0.0	6.00	6.00
GMRT						
4. Level A	136	0.12	0.39	0.0	2.00	2.00
5. Level D	260	0.65	0.48	0.0	1.00	1.00
6. Level F	260	0.85	0.36	0.0	1.00	1.00
Total	656	0.62	0.50	0.0	2.00	2.00
SAT						
7. Primary I	152	0.22	0.41	0.0	1.00	1.00
8. Intermediate I	240	0.57	0.53	0.0	2.00	2.00
9. Intermediate II	256	0.73	0.53	0.0	2.00	2.00
10. High School	260	1.01	0.78	0.0	5.00	5.00
Total	908	0.68	0.65	0.0	5.00	5.00

Table 11

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Dale-Chall Words in the Choices by Test

Test	Number of Choices	Mean number of Non-D-C Words	Standard Deviation	Minimum number of Non-D-C Words	Maximum number of Non-D-C Words	Range
CAT						
1. Lower Primary	38	0.0	0.0	0.0	0.0	0.0
2. Elementary	120	0.32	0.56	0.0	2.00	2.00
3. Advanced	180	1.28	1.03	0.0	6.00	6.00
Total	338	0.80	0.98	0.0	6.00	6.00
GMRT						
4. Level A	136	0.04	0.21	0.0	1.00	1.00
5. Level D	260	0.30	0.46	0.0	1.00	1.00
6. Level F	260	0.65	0.48	0.0	1.00	1.00
Total	656	0.39	0.49	0.0	1.00	1.00
SAT						
7. Primary I	152	0.03	0.16	0.0	1.00	1.00
8. Intermediate I	240	0.30	0.50	0.0	2.00	2.00
9. Intermediate II	256	0.43	0.51	0.0	2.00	2.00
10. High School	260	0.83	0.76	0.0	4.00	4.00
Total	908	0.44	0.62	0.0	4.00	4.00

Table 12

Mean and Standard Deviation for Sentence Length, Spache Ratio, Dale-Chall Ratio; Dale-Chall Raw Scores, Spache Grade Scores, and Dale-Chall Grade Scores in the Reading Selections by Test

Test	Number of Selections	Average Sentence Length		Ratios				Dale-Chall Raw Score		Grade Scores			
				Spache		Dale-Chall				Spache		Dale-Chall	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
CAT													
1. Lower Primary	4	6.46	1.64	5.31	4.54	0.0	0.0	3.96	0.08	2.21	0.26	4.00[a]	0.0
2. Elementary	3	15.04	1.43	28.18	5.81	14.48	4.21	6.67	0.70	5.38[a]	0.51	8.17	2.31
3. Advanced	5	21.29	3.57	30.69	4.03	27.71	8.98	9.07	1.45	6.48[a]	0.69	12.50	3.30
Total	12	14.78	7.08	21.60	12.79	15.17	13.71	6.76	2.48	4.78[a]	2.02	8.58	4.43
GMRT													
4. Level A	16	8.95	2.35	12.22	9.60	1.74	4.02	4.35	0.67	3.15	1.00	4.31[a]	0.93
5. Level D	21	15.27	3.54	26.70	8.90	12.77	9.64	6.41	1.59	5.29[a]	1.01	7.52	3.01
6. Level F	21	23.81	9.28	35.97	5.94	28.17	7.17	9.27	1.04	7.29[a]	1.40	13.36	1.99
Total	58	16.62	8.50	26.06	12.43	15.30	13.04	6.88	2.31	5.42[a]	2.01	8.75	4.32
SAT													
7. Primary I	33	5.21	1.02	3.63	4.86	0.19	1.09	3.92	0.18	1.89	0.49	4.00[a]	0.0
8. Intermediate I	24	16.68	6.60	22.04	7.36	11.83	8.43	6.33	1.37	5.09[a]	1.05	7.38	2.76
9. Intermediate II	25	18.93	7.70	26.79	8.40	17.03	8.57	7.27	1.43	5.81[a]	1.19	9.04	2.39
10. High School	13	25.24	9.82	31.20	6.10	24.74	8.98	8.80	1.44	7.08[a]	1.62	12.35	2.97
Total	95	14.46	9.55	18.15	12.85	10.92	11.13	6.08	2.07	4.44[a]	2.21	7.32	3.55

[a] Invalid as grade scores. As noted above, the Spache formula was intended by its author only for grades 1 to 3. The Dale-Chall formula was intended by its authors only for grades 4 and above. Consequently, the Spache grade scores for the intermediate and advanced test levels and the Dale-Chall grade for the lowest test level only demonstrate the relationships existing among the test levels and between the readability formulae.

Table 13

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Words in Reading Selections by Test

Test	N[a]	Mean[b] number of Words	Standard Deviation	Minimum number of Words	Maximum number of Words	Range
CAT						
1. Lower Primary	15	15.53	12.05	0.0	28.00	28.00
2. Elementary	30	201.60	43.31	156.00	267.00	111.00
3. Advanced	45	468.56	145.82	229.00	637.00	408.00
GMRT						
4. Level A	34	9.59	11.85	0.0	35.00	35.00
5. Level D	52	44.10	9.85	28.00	71.00	43.00
6. Level F	52	58.92	20.82	34.00	109.00	75.00
SAT						
7. Primary I	38	18.68	8.83	8.00	42.00	34.00
8. Intermediate I	60	77.30	40.37	13.00	161.00	148.00
9. Intermediate II	64	78.42	30.92	12.00	127.00	115.00
10. High School	65	168.95	104.07	39.00	383.00	344.00

[a] N indicates that selection scores were weighted by the number of questions that went with a given selection, e.g., if a selection had 4 questions the scores for the selection were included 4 times.

[b] Some questions had no selections, just questions and choices, e.g.,"Which is the big tree (GMRT, Level A, item 2)?" The question was followed by 4 picture choices, one of which was a big tree. In weighting scores "0"s were included as selection data for such items.

Table 14

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Sentences in the Reading Selections by Tests

Test	N[a]	Mean[b] number of Sentences	Standard Deviation	Minimum number of Sentences	Maximum number of Sentences	Range
CAT						
1. Lower Primary	15	2.60	1.98	0.0	5.00	5.00
2. Elementary	30	12.70	1.68	11.00	16.00	5.00
3. Advanced	45	23.78	9.96	12.00	36.00	24.00
GMRT						
4. Level A	34	1.12	1.41	0.0	4.00	4.00
5. Level D	52	2.96	0.68	2.00	4.00	2.00
6. Level F	52	2.62	0.89	1.00	4.00	3.00
SAT						
7. Primary I	38	3.42	1.11	2.00	6.00	4.00
8. Intermediate I	60	4.67	2.48	2.00	10.00	8.00
9. Intermediate II	64	4.11	1.91	1.00	9.00	8.00
10. High School	65	6.12	2.91	2.00	12.00	10.00

[a] N indicates that selection scores were weighted by the number of questions that went with a given selection, e.g., if a selection had 4 questions the scores for the selection were included 4 times.

[b] Some questions had no selections, just questions and choices, e.g.,"Which is the big tree (GMRT, Level A, item 2)?" The question was followed by 4 picture choices, one of which was a big tree. In weighting scores "0"s were included as selection data for such items.

Table 15

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Spache Words in the Reading Selections by Test

Test	N[a]	Mean[b] number of Non-Spache Words	Standard Deviation	Minimum number of Non-Spache Words	Maximum number of Non-Spache Words	Range
CAT						
1. Lower Primary	15	1.13	0.99	0.0	2.00	2.00
2. Elementary	30	57.43	13.87	36.00	72.00	36.00
3. Advanced	45	135.84	34.95	81.00	182.00	101.00
GMRT						
4. Level A	34	1.03	1.57	0.0	5.00	5.00
5. Level D	52	12.08	4.67	4.00	21.00	17.00
6. Level F	52	20.60	6.37	11.00	33.00	22.00
SAT						
7. Primary I	38	0.76	0.94	0.0	3.00	3.00
8. Intermediate I	60	17.40	12.18	2.00	48.00	46.00
9. Intermediate II	64	21.45	11.37	6.00	41.00	35.00
10. High School	65	53.20	31.09	8.00	100.00	92.00

[a] N indicates that selection scores were weighted by the number of questions that went with a given selection, e.g., if a selection had 4 questions the scores for the selection were included 4 times.

[b] Some questions had no selections, just questions and choices, e.g.,"Which is the big tree (GMRT, Level A, item 2)? "The question was followed by 4 picture choices, one of which was a big tree. In weighting scores "0"s were included as selection data for such items.

Table 16

Mean, Standard Deviation, Minimum, Maximum and Range of the Number of Non-Dale-Chall Words in Reading Selections by Test

Test	N[a]	Mean[b] number of Non-D-C Words	Standard Deviation	Minimum number of Non-D-C Words	Maximum number of Non D-C Words	Range
CAT						
1. Lower Primary	15	0.0	0.0	0.0	0.0	0.0
2. Elementary	30	32.43	12.48	15.00	45.00	30.00
3. Advanced	45	117.11	30.51	70.00	154.00	84.00
GMRT						
4. Level A	34	0.18	0.72	0.0	4.00	4.00
5. Level D	52	6.14	5.11	0.0	21.00	21.00
6. Level F	52	16.36	5.56	7.00	27.00	20.00
SAT						
7. Primary I	38	0.05	0.23	0.0	1.00	1.00
8. Intermediate I	60	10.98	8.86	0.0	26.00	26.00
9. Intermediate II	64	13.30	6.46	2.00	22.00	20.00
10. High School	65	40.75	28.62	4.00	106.00	102.00

[a] N indicates that selection scores were weighted by the number of questions that went with a given selection, e.g., if a selection had 4 questions the scores for the selection were included 4 times.

[b] Some questions had no selections, just questions and choices, e.g.,"Which is the big tree (GMRT, Level A, item 2)?" The question was followed by 4 picture choices, one of which was a big tree. In weighting scores "0"s were included as selection data for such items.

Table 17

Mean, Standard Deviation, Minimum, Maximum, and Range for the Spache and Dale-Chall Ratios in the Reading Selections by Test

Test	N[a]	Spache Ratio[b]					Dale-Chall Ratio[b]				
		Mean	S.D.	Minimum	Maximum	Range	Mean	S.D.	Minimum	Maximum	Range
CAT											
1. Lower Primary	15	4.95	4.57	0.0	10.53	10.53	0.0	0.0	0.0	0.0	0.0
2. Elementary	30	28.48	4.30	23.08	34.50	11.42	15.56	3.74	9.62	18.88	9.26
3. Advanced	45	29.78	3.43	25.71	35.37	9.66	26.35	7.34	18.97	37.54	18.57
GMRT											
4. Level A	34	5.75	8.96	0.0	33.33	33.33	0.82	2.85	0.0	12.12	12.12
5. Level D	52	27.20	8.45	8.51	40.54	32.03	13.51	9.96	0.0	38.89	38.89
6. Level F	52	35.84	5.96	22.95	45.83	22.88	28.33	7.10	17.78	45.24	27.46
SAT											
7. Primary I	38	3.66	4.58	0.0	15.00	15.00	0.33	1.41	0.0	6.25	6.25
8. Intermediate I	60	22.02	6.45	12.50	38.46	25.96	12.88	7.64	0.0	31.25	31.25
9. Intermediate II	64	27.27	8.00	13.64	50.00	36.36	17.14	7.70	5.7	48.57	42.86
10. High School	65	31.59	5.61	20.51	39.39	18.88	24.21	8.22	10.26	40.82	30.56

[a] N indicates that selection scores were weighted by the number of questions with a given selection, e.g. if a selection had 4 questions with it the scores for the selection were included 4 times (for further clarification see page ;

[b] Some items had no selections, just questions and choices, e.g. "Which is the big tree (GMRT, Level A, item 2)?" The question was followed by 4 choices, one of which was a big tree. In weighting scores, "0"s were included as selection data for such items.

Table 18

Mean, Standard Deviation, Minimum, Maximum and Range for the Dale-Chall Raw Scores in the Reading Selections by Test

Test	N[a]	Mean[b] Raw Score	Standard Deviation	Minimum Raw Score	Maximum Raw Score	Range
CAT						
1. Lower Primary	15	2.62	1.92	0.0	4.06	4.06
2. Elementary	30	6.88	0.65	5.86	7.51	1.65
3. Advanced	45	8.82	1.22	7.85	10.83	2.98
GMRT						
4. Level A	34	2.05	2.25	0.0	6.10	6.10
5. Level D	52	6.53	1.63	4.00	10.67	6.67
6. Level F	52	9.31	1.02	7.45	11.47	4.02
SAT						
7. Primary I	38	3.95	0.23	3.80	4.89	1.09
8. Intermediate I	60	6.52	1.21	4.03	9.21	5.18
9. Intermediate II	64	7.36	1.28	5.12	12.17	7.05
10. High School	65	8.82	1.30	6.22	11.30	5.08

[a] N indicates that selection scores were weighted by the number of questions that went with a given selection, e.g., if a selection had 4 questions the scores for the selection were included 4 times.

[b] Some questions had no selections, just questions and choices, e.g.,"Which is the big tree (GMRT, Level A, item 2)?" The question was followed by 4 picture choices, one of which was a big tree. In weighting scores "0"s were included as selection data for such items.

Table 19

Mean, Standard Deviation, Minimum, Maximum and Range for the Spache and Dale-Chall Ratios in the Questions by Test

Test	N	Spache Ratio					Dale-Chall Ratio				
		Mean	S.D.	Minimum	Maximum	Range	Mean	S.D.	Minimum	Maximum	Range
CAT											
1. Lower Primary	15	3.93	4.52	0.0	12.50	12.50	0.95	3.69	0.0	14.29	14.29
2. Elementary	30	24.76	11.99	0.0	50.00	50.00	19.05	14.01	0.0	55.56	55.56
3. Advanced	45	34.88	13.98	12.12	75.00	62.88	29.05	12.88	0.0	66.67	66.67
GMRT											
4. Level A	34	11.71	14.62	0.0	37.50	37.50	0.0	0.0	0.0	0.0	0.0
5. Level D	52	26.89	13.44	0.0	56.25	56.25	12.63	12.28	0.0	45.45	45.45
6. Level F	52	36.17	10.70	7.89	64.29	56.40	28.21	11.14	0.0	50.00	50.00
SAT											
7. Primary I	38	3.46	9.44	0.0	40.00	40.00	0.0	0.0	0.0	0.0	0.0
8. Intermeiate I	60	23.00	11.01	0.0	50.00	50.00	12.23	11.00	0.0	46.15	46.15
9. Intermediate II	64	28.89	12.65	0.0	66.67	66.67	19.09	12.71	0.0	50.00	50.00
10. High School	65	35.99	12.05	12.50	71.43	58.93	27.84	11.63	8.33	62.50	54.17

Table 20

Mean, Standard Deviation, Minimum, Maximum and Range for the Spache and Dale-Chall Ratios in the Choices by Test

Test	N[a]	Spache Ratio					Dale-Chall Ratio				
		Mean	S.D.	Minimum	Maximum	Range	Mean	S.D.	Minimum	Maximum	Range
CAT											
1. Lower Primary	15	10.00	26.58	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
2. Elementary	30	40.87	29.49	0.0	100.0	100.0	14.07	16.83	0.0	57.14	57.14
3. Advanced	45	50.97	28.40	0.0	100.0	100.0	49.14	30.53	0.0	100.0	100.0
GMRT											
4. Level A	34	7.54	16.95	0.0	60.00	60.00	2.33	6.62	0.0	25.00	25.00
5. Level D	52	64.29	25.81	20.00	100.0	80.00	30.06	27.98	0.0	100.0	100.0
6. Level F	52	84.99	21.64	0.0	100.0	100.0	64.61	22.96	0.0	100.0	100.0
SAT											
7. Primary I	38	21.58	27.36	0.0	100.0	100.0	2.63	7.78	0.0	25.00	25.00
8. Intermediate I	60	47.44	28.94	0.0	100.0	100.0	24.08	22.96	0.0	75.00	75.00
9. Intermediate II	64	60.37	29.71	0.0	100.0	100.0	35.17	28.27	0.0	100.0	100.0
10. High School	65	67.30	33.51	0.0	100.0	100.0	54.63	35.84	0.0	100.0	100.0

[a] The number of choices corresponds to the number of "sets" of choices that accompanied one test question, i.e. means and standard deviations were computed on the sum of a given score, e.g. Spache ratio, for the four or five choices with a given question.

Table 21

Mean, Standard Deviation, Minimum, Maximum and Range of Difficulty Scores by Grades and Tests

Test	Grade								
	1	2	4	5	6	9	10	11	12
CAT									
1. Number of Scores	15	15	30	30	30	45	45	--	45
2. Mean	26.0	43.4	38.8	48.4	59.8	44.9	54.1	--	64.2
3. S.D.	13.2	12.8	21.9	19.4	19.9	13.8	15.7	--	16.6
4. Minimum	8.0	25.5	12.3	18.5	23.8	19.1	20.7	--	20.4
5. Maximum	59.1	71.6	81.5	82.4	92.4	78.2	87.5	--	88.3
6. Range	51.1	46.1	69.2	63.9	68.6	59.1	66.8	--	67.9
GMRT									
1. Number of Scores	34	--	52	--	52		52[a]		
2. Mean	56.7	--	53.8	--	71.7		48.5		
3. S.D.	21.9	--	23.3	--	18.9		20.8		
4. Minimum	5.3	--	2.2	--	22.0		7.5		
5. Maximum	94.3	--	92.0	--	95.7		85.5		
6. Range	89.0	--	89.8	--	73.7		78.0		
SAT									
1. Number of Scores	38	--	60	64	64	65	65	65	65
2. Mean	43.3	--	51.2	50.3	59.8	43.8	53.1	57.9	61.3
3. S.D.	18.5	--	20.5	18.0	17.9	16.5	17.4	17.5	16.8
4. Minimum	17.0	--	11.0	13.0	17.0	15.0	16.0	19.0	21.0
5. Maximum	86.0	--	88.0	84.0	89.0	75.0	86.0	89.0	92.0
6. Range	69.0	--	77.0	71.0	72.0	60.0	70.0	70.0	71.0

Note: A hyphen means that no data were provided by the test-authors.

[a] These figures represent a composite score for grades 10, 11, and 12.

Table 22

Miminum, Maximum and Range for the Number of Words, Non-Spache and Non-Dale-Chall Words in the Reading Selection, Question and Choices by Test Level[a]

Test Levels	N	Number of Words			Non-Spache Words			Non-Dale-Chall Words		
		Minimum	Maximum	Range	Minimum	Maximum	Range	Minimum	Maximum	Range
Lowest (Grades 1-2)										
Selection	53	7	42	35	0	5	5	0	4	4
Question	87	0	22	22	0	3	3	0	1	1
Choice	326	0	7	7	0	2	2	0	1	1
Intermediate (Grades 4-6)										
Selection	73	12	267	255	2	72	70	0	45	45
Question	206	3	49	46	0	13	13	0	10	10
Choice	876	1	6	5	0	2	2	0	2	2
Advanced (Grades 9-12)										
Selection	39	34	637	603	8	182	174	4	154	150
Question	162	4	47	43	1	20	19	0	13	13
Choice	700	1	19	18	0	6	6	0	6	6

[a] Data for the three test batteries (CAT, GMRT, SAT) are combined.

Table 23

Correlations of Readability and Difficulty Scores
for the CAT Lower Primary

Selection	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. No. of words																				
2. No. of sentences	95																			
3. Avg. sentence length	93	84																		
4. No. non-Spache words	84	90	69																	
5. No. non-Dale-Chall words	—	—	—	—																
6. Spache ratio	71	84	58	96	—															
7. Dale-Chall ratio	—	—	—	—	—	—														
8. Spache Grade Score	94	95	92	89	—	85	—													
9. Dale-Chall Grade Score	94	95	94	84	—	79	—	99												
Question																				
10. No. of words	61	61	66	24	—	21	—	56	66											
11. No. non-Spache words	15	12	09	29	—	24	—	15	09	-06										
12. No. non-Dale-Chall words	-36	-36	-36	-32	—	-30	—	-37	-38	-39	-25									
13. Spache ratio	03	-01	-03	15	—	10	—	01	+04	-15	96	-24								
14. Dale-Chall ratio	-36	-36	-36	-32	—	-30	—	-37	-38	-39	-25	100	-24							
Choice																				
15. No. of words	45	42	47	-01	—	-07	—	32	43	90	-30	-20	-33	-20						
16. No. non-Spache words	13	02	10	10	—	-02	—	03	00	-16	53	-13	62	-13	-11					
17. No. non-Dale-Chall words	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—				
18. Spache ratio	-21	-26	-22	-19	—	-23	—	-26	-28	-30	42	-10	63	-10	-20	78	—			
19. Dale-Chall ratio	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—		
Difficulty Score																				
20. Grade 1	-67	-71	-64	-53	—	-51	—	-68	-71	-70	07	17	19	17	-58	05	—	33	—	
21. Grade 2	-29	-31	-33	-23	—	-25	—	-34	-36	-50	-03	34	08	34	-33	03	—	29	—	78

Note 1: A hyphen means that one score was constant throughout all items of the test and therefore the correlation was meaningless.

Note 2: The number of items in this test was 15. The critical correlation for 14 degrees of freedom at two significance is: P = .05, r = 50; P = .01, r = 62.

Table 24

Correlations of Readability and Difficulty Scores
for the CAT Elementary

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Selection																					
1. No. of words																					
2. No. of sentences	92																				
3. Avg. sentence length	87	61																			
4. No. non-Spache words	86	81	74																		
5. No. non-Dale-Chall words	93	81	88	97																	
6. Spache ratio	02	07	-03	53	34																
7. Dale-Chall ratio	68	56	70	94	90	69															
8. Spache Grade Score	49	40	53	86	77	84	97														
9. Dale-Chall Grade Score	65	62	56	95	86	77	97	96													
Question																					
10. No. of words	74	78	51	51	56	-22	-27	10	29												
11. No. non-Spache words	72	78	48	59	60	-03	38	24	42	93											
12. No. non-Dale-Chall words	80	90	47	59	61	-13	32	15	38	88	87										
13. Spache ratio	-13	-05	-18	16	03	53	25	35	32	-21	14	-06									
14. Dale-Chall ratio	29	26	27	30	31	10	28	24	27	-02	08	36	26								
Choice																					
15. No. of words	-49	-46	-40	-35	-41	11	-24	-13	-22	-49	-49	-51	05	-17							
16. No. non-Spache words	-45	-39	-42	-19	30	38	-05	09	00	-63	-50	-46	37	14	37						
17. No. non-Dale-Chall words	-14	-20	-02	09	-04	40	25	33	23	-54	-37	-28	46	48	40	49					
18. Spache ratio	-11	-04	-16	00	-06	18	02	07	06	-20	-12	-04	18	18	-30	69	08				
19. Dale-Chall ratio	-14	-22	-02	09	05	38	25	33	22	-61	-47	-32	42	55	10	49	84	30			
Difficulty Score																					
20. Grade 4	-83	-74	-77	-90	-91	-36	-84	-73	-81	-58	-68	-63	-18	-37	38	24	-14	04	-07		
21. Grade 5	-80	-71	-75	-87	-68	-35	-81	-71	-79	-60	-69	-61	-14	-28	37	27	-11	07	-00	98	
22. Grade 6	-82	-73	-74	-87	-88	-34	-80	-70	-78	-60	-67	-62	-13	-26	33	27	-11	10	01	95	98

Note: The number of items in this test was 30. The critical correlation for 29 degrees of freedom at two significance levels is: P = .05, r = 36; P = .01, r = 46.

Table 25

Correlations of Readability and Difficulty Scores
for the CAT Advanced

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Selection																					
1. No. of words																					
2. No. of sentences	98																				
3. Avg. sentence length	-68	-80																			
4. No. non-Spache words	97	96	-62																		
5. No. non-Dale-Chall words	74	76	-50	-17																	
6. Spache ratio	-76	-70	47	-57	-17																
7. Dale-Chall ratio	-60	-53	31	-43	08	90															
8. Spache Grade Score	-82	-68	92	-70	-43	79	62														
9. Dale-Chall Grade Score	-58	-50	29	-39	11	92	99	62													
Question																					
10. No. of words	15	11	01	13	16	-14	-04	-06	-06												
11. No. non-Spache words	-15	-19	22	-16	-17	05	01	18	00	48											
12. No. non-Dale-Chall words	-17	-16	08	-17	-09	15	16	12	15	43	50										
13. Spache ratio	-35	-34	20	-37	-45	17	-02	22	-01	-40	49	-08									
14. Dale-Chall ratio	-34	-27	-02	-33	-23	31	28	13	28	-32	-03	57	08								
Choice																					
15. No. of words	18	17	-11	17	24	-13	03	-14	01	60	23	29	-29	-23							
16. No. non-Spache words	-03	01	-01	-03	-04	-16	-07	-08	-10	54	51	53	04	00	72						
17. No. non-Dale-Chall words	-02	-02	-05	-05	-02	-03	03	-05	02	50	28	65	-21	19	66	71					
18. Spache ratio	-23	-21	05	-28	-39	04	-10	05	-10	-37	06	01	44	33	-56	-01	-19				
19. Dale-Chall ratio	-18	-16	03	-19	-25	10	-02	07	-01	-38	-15	06	14	44	-62	-36	02	61			
Difficulty Score																					
20. Grade 9	39	35	-23	28	11	-53	-45	-41	-46	-05	08	-16	12	-21	16	20	05	-02	-06		
21. Grade 10	33	30	-20	23	09	-45	-36	-35	-38	-11	08	-15	17	-17	14	20	03	03	-05	96	
22. Grade 12	29	26	-15	20	04	-44	-38	-31	-40	-18	02	-19	18	-13	11	16	01	03	-01	89	93

Note: The number of items in this test was 45. The critical correlation for 44 degrees of freedom at two signficance levels is: P = .05, r = 29; P = .01, r = 38.

Table 26

Correlations of Readability and Difficulty Scores for the GMRT Level A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Selection																			
1. No. of words																			
2. No. of sentences	95																		
3. Avg. sentence length	86	74																	
4. No. non-Spache words	77	62	81																
5. No. non-Dale-Chall words	38	25	34	51															
6. Spache ratio	46	36	75	80	29														
7. Dale-Chall ratio	30	19	37	46	88	39													
8. Spache Grade Score	76	66	95	85	33	91	39												
9. Dale-Chall Grade Score	87	82	93	76	51	70	53	91											
Question																			
10. No. of words	-40	-33	-57	-51	-33	-60	-38	-63	-59										
11. No. non-Spache words	-43	-42	-46	-35	-13	-34	-15	-45	-47	39									
12. No. non-Dale-Chall words	--	--	--	--	--	--	--	--	--	--	--								
13. Spache ratio	28	20	19	24	29	02	16	12	20	-42	39	--							
14. Dale-Chall ratio	--	--	--	--	--	--	--	--	--	--	--	--	--						
Choice																			
15. No. of words	-20	-16	-24	-21	-12	-22	-13	-25	-23	57	55	--	04	--					
16. No. non-Spache words	-15	-12	-21	-17	-10	-19	-11	-21	-19	44	55	--	11	--	87				
17. No. non-Dale-Chall words	-07	-03	-13	-12	-09	-14	-10	-13	-11	37	33	--	02	--	90	95			
18. Spache ratio	-10	-04	-19	-15	-11	-18	-13	-19	-17	44	46	--	03	--	77	87	80		
19. Dale-Chall ratio	06	13	-05	-05	-09	-09	-10	-06	-01	32	21	--	-06	--	72	85	93	83	
Difficulty Score																			
20. Grade 1	-61	-64	-54	-43	-12	-31	-06	-50	-56	-01	03	--	-23	--	-35	-33	-34	-40	-39

Note 1: A hyphen means that one score was constant throughout all items of the test and therefore the correlation was meaningless.

Note 2: The number of items in this test was 34. The critical correlation for 33 degrees of freedom at two levels of significance is: P = .05, r = 33; P = .01, r = 43.

Table 27

Correlations of Readability and Difficulty Scores

for the <u>GMRT</u> Level D

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
<u>Selection</u>																				
1. No. of words																				
2. No. of sentences	47																			
3. Avg. sentence length	36	-63																		
4. No. non-Spache words	65	11	40																	
5. No. non-Dale-Chall words	42	-04	36	78																
6. Spache ratio	10	-22	26	82	66															
7. Dale-Chall ratio	18	-21	33	70	95	74														
8. Spache Grade Score	25	-47	69	81	67	88	72													
9. Dale-Chall Grade Score	27	-19	40	74	96	73	98	75												
<u>Question</u>																				
10. No. of words	23	-39	60	31	34	23	32	46	35											
11. No. non-Spache words	20	-27	42	64	68	66	70	71	72	72										
12. No. non-Dale-Chall words	17	-21	34	59	83	61	86	63	84	49	83									
13. Spache ratio	13	-21	26	68	66	79	74	72	72	29	81	70								
14. Dale-Chall ratio	13	-18	26	60	82	68	89	64	85	23	66	91	75							
<u>Choice</u>																				
15. No. of words	-10	01	-13	06	13	18	23	07	17	-13	02	-01	20	06						
16. No. non-Spache words	-06	-25	22	26	26	36	30	37	29	17	28	25	33	28	19					
17. No. non-Dale-Chall words	05	-28	33	42	44	52	51	56	49	11	36	43	48	52	34	58				
18. Spache ratio	-05	-26	23	26	25	35	28	37	28	18	28	26	31	27	10	100	56			
19. Dale-Chall ratio	05	-29	35	43	44	52	51	56	49	12	37	45	47	53	27	58	100	56		
<u>Difficulty Score</u>																				
20. Grade 4	-29	07	-33	-53	-54	-48	-55	-52	-60	-11	-42	-46	-54	-55	-01	-33	-53	-33	-54	
21. Grade 6	-30	-00	-28	-48	-56	-41	-56	-45	-62	-07	-36	-44	-45	-51	-08	27	-52	-27	-53	95

Note: The number of items in this test was 52. The critical correlation for 51 degrees of freedom at two significance levels is: P = .05, r = 27; P = .01, r = 35.

Table 28

Correlations of Raadability and Difficulty Scores
for the GMRT Level F

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Selection																			
1. No. of words																			
2. No. of sentences	53																		
3. Avg. sentence length	34	-56																	
4. No. non-Spache words	87	53	25																
5. No. non-Dale-Chall words	80	60	06	80															
6. Spache ratio	-43	-22	-11	06	-17														
7. Dale-Chall ratio	-22	14	-42	-06	38	36													
8. Spache Grade Score	16	-63	92	27	-01	29	-27												
9. Dale-Chall Grade Score	03	-06	07	15	52	31	83	19											
Question																			
10. No. of words	27	-42	77	22	-01	-06	-42	71	-06										
11. No. non-Spache words	12	-43	69	26	-02	32	-22	80	14	76									
12. No. non-Dale-Chall words	14	-33	52	23	28	18	23	57	51	52	63								
13. Spache ratio	-43	-05	-07	11	01	60	31	17	29	-26	39	17							
14. Dale-Chall ratio	-22	09	-24	03	32	28	73	-12	63	-43	-11	47	46						
Choice																			
15. No. of words	—	—	—	—	—	—	—	—	—	—	—	—	—	—					
16. No. non-Spache words	19	24	03	21	22	-02	06	02	17	15	22	16	01	00	—				
17. No. non-Dale-Chall words	37	26	06	33	36	-17	00	-01	11	13	06	15	-18	03	—	65			
18. Spache ratio	19	24	03	21	22	-02	06	03	17	15	22	16	01	00	—	100	65		
19. Dale-Chall ratio	37	26	06	33	36	-17	00	-01	11	13	06	15	-18	03	—	65	100	65	
Difficulty Score																			
20. Grades 10, 11, 12	-33	-04	-27	-41	-42	-05	-10	-28	-29	-16	-14	-33	-02	-21	—	-06	-18	-06	-18

Note 1: A hyphen means that one score was constant throughout all items of the test and therefore the correlation was meaningless.

Note 2: The number of items in this test was 52. The critical correlation for 51 degrees of freedom at two significance levels is: P = .05, r = 27; P = .01, r = 35.

Table 29

Correlations of Readability and Difficulty Scores for the SAT Primary 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Selection																			
1. No. of words																			
2. No. of sentences	92																		
3. Avg. sentence length	68	36																	
4. No. non-Spache words	56	46	49																
5. No. non-Dale-Chall words	-07	-09	-00	06															
6. Spache ratio	20	11	32	90	14														
7. Dale-Chall ratio	-07	-09	-00	06	100	14													
8. Spache Grade Score	39	21	59	92	11	95	11												
9. Dale-Chall Grade Score	—	—	—	—	—	—	—	—											
Question																			
10. No. of words	62	46	57	31	02	06	02	23	—										
11. No. non-Spache words	13	14	10	29	-09	34	-09	32	—	-06									
12. No. non-Dale-Chall words	—	—	—	—	—	—	—	—	—	—	—								
13. Spache ratio	12	14	09	29	-09	35	-09	32	—	-09	99	—							
14. Dale-Chall ratio	—	—	—	—	—	—	—	—	—	—	—	—	—						
Choice																			
15. No. of words	01	02	03	06	-06	12	-06	11	—	-24	19	—	23	—					
16. No. non-Spache words	31	27	32	34	-19	28	-19	34	—	11	67	—	62	—	-08				
17. No. non-Dale-Chall words	33	26	30	18	-08	10	-08	18	—	15	47	—	39	—	-08	52			
18. Spache ratio	32	27	32	33	-19	27	-19	33	—	12	66	—	61	—	-10	100	52		
19. Dale-Chall ratio	33	26	30	18	-08	10	-08	18	—	15	47	—	39	—	-08	52	100	52	
Difficulty Score																			
20. Grade 1	-46	-39	-44	-42	-13	-37	-13	-45	—	-22	-32	—	-30	—	-16	-44	-32	-44	-32

Note 1: A hyphen means that one score was constant throughout all items of the test and therefore the correlation was meaningless.

Note 2: The number of items in this test was 38. The critical correlation for 37 degrees of freedom at two significance levels is: P = .05, r = 32; P = .01, r = 41.

Table 30

Correlations of Readability and Difficulty Scores
for the SAT Intermediate I

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Selection																			
1. No. of words																			
2. No. of sentences	90																		
3. Avg. sentence length	22	-19																	
4. No. non-Spache words	91	90	08																
5. No. non-Dale-Chall words	84	87	-00	87															
6. Spache ratio	15	23	-10	49	33														
7. Dale-Chall ratio	34	43	-10	40	74	34													
8. Spache Grade Score	28	-03	80	36	20	51	11												
9. Dale-Chall Grade Score	33	39	-00	41	73	35	97	21											
Question																			
10. No. of words	-12	-41	74	-30	-23	-33	-13	45	-07										
11. No. non-Spache words	-15	-27	32	-09	-05	21	12	41	17	52									
12. No. non-Dale-Chall words	19	08	30	12	38	00	55	26	55	47	61								
13. Spache ratio	13	22	-14	33	30	50	22	17	27	-18	63	28							
14. Dale-Chall ratio	39	44	-03	41	66	18	71	08	68	-02	36	81	46						
Choice																			
15. No. of words	25	30	-02	24	31	03	22	-00	21	-13	-24	-11	-07	01					
16. No. non-Spache words	31	17	29	26	19	02	-04	26	00	06	-00	-07	06	-04	20				
17. No. non-Dale-Chall words	10	13	01	02	31	-12	50	-07	48	06	-03	32	-08	39	43	26			
18. Spache ratio	12	-04	32	08	01	-02	-11	26	-06	18	12	06	03	01	-33	83	13		
19. Dale-Chall ratio	07	04	12	-03	22	-13	40	03	40	21	17	45	02	42	-08	29	81	39	
Difficulty Score																			
20. Grade 4	-22	-23	-07	-21	-22	-10	-13	-12	-07	-09	-09	-24	-17	-29	-19	-09	-17	-02	-12

Note: The number of items in this test was 60. The critical correlation for 59 degrees of freedom at two significance levels is: P = .05, r = 25; P = .01, r = 33.

Table 31

Correlations of Readability and Difficulty Scores
for the SAT Intermediate II

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Selection																				
1. No. of words																				
2. No. of sentences	64																			
3. Avg. sentence length	45	-38																		
4. No. non-Spache words	84	63	27																	
5. No. non-Dale-Chall words	77	49	34	68																
6. Spache ratio	03	11	-12	53	17															
7. Dale-Chall ratio	-06	-05	-01	04	53	33														
8. Spache Grade Score	42	-29	85	52	40	42	17													
9. Dale-Chall Grade Score	10	-20	37	06	63	13	89	41												
Question																				
10. No. of words	-04	-36	41	-28	-09	-38	05	17	12											
11. No. non-Spache words	-11	-24	21	-14	02	-01	18	19	23	67										
12. No. non-Dale-Chall words	08	-01	15	-04	28	-06	46	11	45	56	61									
13. Spache ratio	-01	26	-26	13	19	33	30	-08	16	-33	34	15								
14. Dale-Chall ratio	16	41	-23	23	36	26	50	-08	32	-25	09	60	55							
Choice																				
15. No. of words	20	-00	25	27	12	20	-05	34	07	-00	-02	04	-11	-01						
16. No. non-Spache words	08	32	-27	25	08	31	05	-08	-11	-31	-05	07	23	35	07					
17. No. non-Dale-Chall words	24	37	-10	31	41	23	33	03	24	-13	-05	17	11	37	10	44				
18. Spache ratio	-16	15	-37	-04	-05	13	12	-27	-07	-31	-09	-00	19	29	-46	77	26			
19. Dale-Chall ratio	08	29	-19	11	33	09	39	-13	27	-09	-05	18	11	37	-25	31	89	43		
Difficulty Score																				
20. Grade 5	-22	07	-40	-26	-33	-16	-25	-45	-41	-09	-19	-15	-17	-09	-24	05	-13	20	-05	
21. Grade 6	-24	06	-41	-27	-34	-15	-22	-45	-38	-06	-16	-09	-16	-03	-23	04	-13	18	-07	97

Note: The number of items in this test was 64. The critical correlation for 63 degrees of freedom at two significance levels is: P = .05, r = 24; P = .01, r = 32.

Table 32

Correlations of Readability and Difficulty Scores for the SAT High School

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Selection																						
1. No. of words																						
2. No. of sentences	80																					
3. Avg. sentence length	61	05																				
4. No. non-Spache words	96	69	72																			
5. No. non-Dale-Chall words	90	84	36	85																		
6. Spache ratio	-03	-25	31	21	06																	
7. Dale-Chall ratio	-02	11	-21	03	35	53																
8. Spache Grade Score	53	-02	96	69	33	55	-04															
9. Dale-Chall Grade Score	21	14	14	30	50	60	91	29														
Question																						
10. No. of words	-22	-27	-13	-29	-09	06	27	-10	15													
11. No. non-Spache words	-14	-22	-05	-17	01	24	35	02	26	83												
12. No. non-Dale-Chall words	-15	-16	-15	-18	11	27	64	-06	52	78	85											
13. Spache ratio	17	06	20	23	16	28	07	25	14	-27	25	07										
14. Dale-Chall ratio	07	14	-11	10	32	30	61	-01	61	-10	18	49	45									
Choice																						
15. No. of words	30	32	08	26	25	-10	-13	04	-08	00	-10	-11	-10	-04								
16. No. non-Spache words	22	21	06	17	21	-15	-02	01	02	-07	-15	-03	-12	07	66							
17. No. non-Dale-Chall words	24	38	-17	16	45	-13	35	-19	34	20	10	27	-19	29	45	43						
18. Spache ratio	-41	-41	-15	-40	-35	-03	12	-14	04	06	04	17	-10	08	-58	08	-20					
19. Dale-Chall ratio	-38	-17	-48	-42	[illegible]	-07	41	-44	24	30	22	40	-27	22	-43	-18	41	48				
Difficulty Score																						
20. Grade 9	-41	-30	-29	-38	-29	04	18	-24	11	01	-07	-04	-14	-04	-20	-18	-07	01	11			
21. Grade 10	-42	-30	-30	-38	-30	05	17	-24	10	-00	-08	-06	-12	-05	-19	-21	-08	-03	09	98		
22. Grade 11	-44	-32	-27	-37	-31	13	22	-20	15	02	-05	-03	-11	-03	-22	-23	-08	00	11	95	97	
23. Grade 12	-46	-34	-29	-41	-34	10	20	-22	11	06	-01	-01	-10	-04	-21	-22	-08	01	13	92	96	98

Note: The number of items in this test was 65. The critical correlation for 64 degrees of freedom at two significance levels is: P = .05, r = 24; P = .01, r = 32.

Table 33

Percent of Reading Selection in Each Code by Level

Selection Code	Number of Selections[a]	Percent Selections by Level			
		Lowest (1-2)	Intermediate (4-6)	Advanced (9-14)	All Grades
1. Riddle	13	24	--	--	8
2. Story	52	72	16	5	32
3. Language	8	--	7	8	5
4. Math	1	--	1	--	1
5. Social Studies	19	--	19	13	12
6. Social Science	11	--	4	20	7
7. Science	50	4	49	31	30
8. Humanities	11	--	3	23	7

Note: A hyphen means no selections were rated in the category.

[a] A total of 165 selections were analyzed.

Table 34

Percent of Questions in Each Code by Level

Question Code	Number of Questions[a]	Percent Selections by Level			
		Lowest (1-2)	Intermediate (4-6)	Advanced (9-14)	All Grades
1. Recognition	12	7	3	--	3
2. Contextual Paraphrase	117	12	35	21	26
3. Grammatical Paraphrase	36	--	11	9	8
4. Semantic Paraphrase	63	5	10	24	14
5. Definite Concept	9	2	2	2	2
6. Probable Concept	93	37	15	18	20
7. Language Concept	34	2	9	9	8
8. Previous Knowledge	68	9	15	18	15
9. Matching[b]	23	26	--	--	5

Note: A hyphen means no questions were rated in the category.

[a] A total of 455 questions were analyzed.

[b] "Matching" items combined the following: 16% Matching (word-picture), 1% Matching and Recognition, 5% Matching and Contextual Paraphrase, 1% Matching and Grammatical Paraphrase, 1% Matching and Semantic Paraphrase, and 2% Matching and Probable Concept for a total of 26%.

Table 35

Percent of Choices in Each Code by Level

Choice Code	Number of Distractors[a]	Percent Choices by Level			All Levels
		Lowest (Grades 1-2)	Intermediate (Grades 4-6)	Advanced (Grades 9-14)	
1. Other	43	2	4	1	3
2. Grammatical	307	16	20	25	21
3. Associative	39	0	3	4	3
4. Associative-Grammatical	456	33	30	33	32
5. Categorical	1	--	0	1	0
6. Categorical-Grammatical	207	13	15	14	14
7. Textual	9	1	1	--	1
8. Textual-Grammatical	56	5	5	2	4
9. Textual-Associative	12	--	1	1	1
10. Textual-Associative-Grammatical	148	14	12	7	10
11. Textual-Categorical	2	--	0	--	0
12. Textual-Categorical-Grammatical	164	15	8	14	11

Note 1: A hyphen means no choices were rated in the category

Note 2: "0" means that less than 1% of the choices were rated in the category

[a] A total of 1444 distractors were analyzed.

Table 36

Percent Reading Selections in Each Code by Test

Test	Number of Selections	Percent Selections by Test 1Riddle	2Story	3Language	4 Math	5Social Studies	6Social Science	7Science	8Humanities
CAT									
1. Lower Primary	4	—	100	—	—	—	—	—	—
2. Elementary	3	—	—	—	—	33	—	67	—
3. Advanced	5	—	—	—	—	40	20	20	20
Total	12	—	33	—	—	25	8	25	8
GMRT									
4. Level A	16	—	88	—	—	—	—	12	—
5. Level D	21		19	5	—	28	5	43	—
6. Level F	21	—	10	5	—	5	14	33	33
Total	58	—	35	3	—	12	7	31	12
SAT									
7. Primary I	33	39	61	—	—	—	—	—	—
8. Intermediate I	24	—	21	4	4	8	4	50	8
9. Intermediate II	25	—	12	12	—	20	4	52	—
10. High School	13	—	—	15	—	15	31	31	8
Total	95	14	30	6	1	10	6	30	3

Note: A hyphen means no selections were rated in the category.

Table 37

Percent Questions in Each Code by Test

Test	Number of Questions	Percent Questions by Test 1 Recognition	2 Contextual Paraphrase	3 Grammatical Paraphrase	4 Semantic Paraphrase	5 Definite Concept	6 Probable Concept	7 Language Concept	8 Previous Knowledge	9 Matching
CAT										
1. Lower Primary	15	40	27	--	--	7	13	7	7	--
2. Elementary	30	17	27	3	3	13	30	--	7	--
3. Advanced	45	--	29	11	40	2	9	2	7	--
Total	90	12	28	7	21	7	17	2	7	--
GMRT										
4. Level A	34	--	--	--	12	--	18	--	3	68[a]
5. Level D	52	--	15	14	6	--	19	17	29	--
6. Level F	52	--	12	6	12	--	25	17	29	--
Total	138	--	10	7	9	--	21	13	22	17
SAT										
7. Primary I	38	--	16	--	--	3	63	3	16	--
8. Intermediate I	60	--	50	8	10	--	8	8	15	--
9. Intermediate II	64	2	42	14	17	--	11	6	8	--
10. High School	65	--	23	9	22	3	20	6	14	--
Total	227	0	34	9	14	1	22	6	14	--

Note 1: A hyphen means no questions were rated in the category.

Note 2: "0" means that less than 1% of the questions were rated in the category.

[a] "Matching" items included a number of categories as follows: 41% Matching (word-picture), 3% Matching and Recognition, 12% Matching and Contextual Paraphrase, 3% Matching and Grammatical Paraphrase, 3% Matching and Semantic Paraphrase, and 6% Matching and Probable Concept, for a total of 68%.

Table 38

Percent of Choices in Each Code by Test

Test	Number of Distractors[a]	1 Other	2 Grammatical	3 Associative	4 Grammatical Associative	5 Categorical	6 Grammatical Categorical	7 Textual	8 Grammatical Textual	9 Associative Textual	10 Grammatical Associative Textual	11 Categorical Textual	12 Grammatical Categorical Textual
CAT													
1. Lower Primary	20	—	5	—	35	—	25	—	—	—	5	—	30
2. Elementary	90	—	20	—	17	—	40	—	4	—	11	—	8
3. Advanced	135	—	30	—	33	—	13	—	4	—	12	—	10
Total	245	—	24	—	27	—	24	—	4	—	11	—	11
GERT													
4. Level A	102	—	14	—	19	—	4	—	10	—	30	—	24
5. Level D	208	14	33	8	28	—	6	3	3	1	3	0	—
6. Level F	208	2	27	7	38	—	14	—	1	2	6	—	3
Total	518	7	27	6	30	—	9	1	4	1	10	0	6
SAT													
7. Primary I	114	5	19	1	46	—	19	3	1	—	1	—	5
8. Intermediate I	180	—	12	1	33	1	14	—	7	—	18	1	14
9. Intermediate II	192	0	15	0	35	—	15	—	6	3	15	—	10
10. High School	195	1	20	2	28	—	14	—	2	0	4	—	29
Total	681	1	16	1	34	1	15	0	4	1	10	0	16

Note 1: A hyphen means no choices were rated in the category.
Note 2: "0" means that less than .1% of the choices were rated in the category.
[a] Distractors are wrong answer choices, right choices were not included in the ratings.

APPENDIX B

COLLECTION OF ODD SOUNDING SENTENCES

APPENDIX B

COLLECTION OF ODD SOUNDING SENTENCES

1. It is a something. (SAT-P1-1)[1]

2. It is a little. (SAT-P1-1)

3. It can go see. (SAT-P1-2)

4. It can go want. (SAT-P1-2)

5. It can go blue. (SAT-P1-2)

6. It is a pretty. (SAT-P1-5)

7. We are at here. (SAT-P1-6)

8. We are at fun. (SAT-P1-6)

9. His nose was big and sleepy. (SAT-P1-23)

10. Pete is a house. (SAT-P1-35)

11. If smallpox virus should enter the air of a vaccinated child, the substance is there to prevent the virus from doing any damage. (SAT-I1-19)

12. The name of the star Procyon means "before the dog," and it was so named because it rises just in advance of Procyon Sirius. (SAT-I1-24)

13. If, on the other hand, it stands together in a field or park, it spreads out much more, and growth is not so restricted to height. (SAT-I1-30).

14. In spite of the general increase in the cost of real estate, I am sure the looks of his home has gone down. (SAT-I1-54)

Note: Distractors are underlined.

[1] (test battery - test level - question number)

SAT = Stanford Achievement Test, Form X, Paragraph Meaning Subtest
P1= Primary 1   I2= Intermediate 2
I1= Intermediate 1   HS= High School

GMRT =Gates-MacGinitie Reading Test, Form 1, Comprehension Subtest
D= Survey D   F= Survey F

15. The other parts of the spot can still see, and the part which sees nothing leads to the impression that there is a black spot floating in the air. (SAT-I2-13)

16. The other parts of the light can still see, and the part which sees nothing leads to the impression that there is a black spot floating in the air. (SAT-I2-13)

17. In Roman times Latin was unknown by the most important people then living on the face of the earth. (SAT-I2-14)

18. The smaller the space to be occupied by the gas, the greater must be the applied water. (SAT-I2-22)

19. The smaller the space to be occupied by the gas, the greater must be the applied pump. (SAT-I2-22)

20. One should not confuse the number of light waves per second, or the frequency of the air, with the rate at which light is traveling. (SAT-I2-52)

21. The moon also travels around the earth in perihelion. (SAT-I2-56)

22. Good thought, like good reading, demands a sharp precision between what is important and what is unimportant. (SAT-HS-1)

23. Good thought, like good reading, demands a sharp evaluation between what is important and what is unimportant. (SAT-HS-1)

24. But when they are the reverse, one can always form an unfavorable opinion of him, because his first mistakes are in making these opinions. (SAT-HS-16)

25. Study in school is an activity that has as one of its chief natures the mastery of school subjects. (SAT-HS-17)

26. This mastery is observed by grades, diplomas, vocational success, status, and approval from others. (SAT-HS-18)

27. The values of such reinforcements induces the student to undertake and carry out study activities. (SAT-HS-19)

28. This energy is produced, not by blowing apart the heavy elements as in fission, but by focusing of light elements. (SAT-HS-43)

29. The children were very empty. (GMRT-D-1)

30. "There's a good strong wind bellow," said Dave. (GMRT-D-5)

31. "There's a good strong wind belong," said Dave. (GMRT-D-5)

32. "There's a good strong wind yesterday," said Dave. (GMRT-D-5)

33. As it is, they look so much like the surrounding snow that hunters often do not see them until they melt. (GMRT-D-8)

34. As it is, they look so much like the surrounding snow that hunters often do not see them until they aren't. (GMRT-D-8)

35. The porter who makes up the beds on a train has other wise too. (GMRT-D-9)

36. For example, he helps the passengers with their comfortable as they arrive at their destinations. (GMRT-D-10)

37. They do not own the foreshore, that strip of time lying between the high-water and low-water marks. (GMRT-D-13)

38. They do not own the foreshore, that strip of land lying between the high-water and low-water storms. (GMRT-D-14)

39. When flowers, it beats its wings so rapidly that they sound like the hum of a tiny motor. (GMRT-D-15)

40. As one looks down a long, straight road, it seems to grow narrower in the time. (GMRT-D-17)

41. As one looks down a long, straight road, it seems to grow narrower in the turnpike. (GMRT-D-17)

42. Telephone poles give the distance of growing smaller as the eye follows a row of them toward the horizon. (GMRT-D-18)

43. Telephone-poles give the score of growing smaller as the eye follows a row of them toward the horizon. (GMRT-D-18)

44. Telephone poles give the call of growing smaller as the eye follows a row of them toward the horizon. (GMRT-D-18)

45. Telephone poles give the height of growing smaller as the eye follows a row of them toward the horizon. (GMRT-D-18)

46. Prior to this it was thought idea for a man to run a "four-minute" mile. (GMRT-D-19)

47. Then in 1961 Herb Elliott of Austrailia ran the mile in three times, fifty-four and a half seconds. (GMRT-D-20)

48. He bettered Bannister's right by nearly five seconds. (GMRT-D-21)

49. He bettered Bannister's timely by nearly five seconds. (GMRT-D-21)

50. "Turnpike" is one name given to those highways where travelers must pay told. (GMRT-D-22)

51. "Turnpike" is one name given to those highways where travelers must pay roads. (GMRT-D-22)

52. All buildings using the turnpikes go through toll gates and thereby share the cost of good roads. (GMRT-D-23)

53. All necessary using the turnpikes go through toll gates and thereby share the cost of good roads. (GMRT-D-23)

54. All ready using the turnpikes go through toll gates and thereby share the cost of good roads. (GMRT-D-23)

55. All without using the turnpikes go through toll gates and thereby share the cost of good roads. (GMRT-D-23)

56. Jet planes now cover the Atlantic Ocean take only a fraction of the time that Lindbergh took. (GMRT-D-25)

57. Jet planes now enter the Atlantic Ocean take only a fraction of the time that Lindbergh took. (GMRT-D-25)

58. Jet planes now going the Atlantic Ocean take only a fraction of the time that Lindbergh took. (GMRT-D-25)

59. Jet planes now crossing the Atlantic Ocean take only a double of the time that Lindbergh took. (GMRT-D-26)

60. Jet planes now crossing the Atlantic Ocean take only a passing of the time that Lindbergh took. (GMRT-D-26)

61. To receive the money, he must show proper face. (GMRT-D-31)

62. To receive the money, he must show proper own. (GMRT-D-31)

63. If the air ways increases to much more than sixteen pounds per square inch, the whole world seems to be pressing down and trying to suffocate you. (GMRT-D-32)

64. As they paddled in to the lakeshore, they saw the log cut which was to be their headquarters for the trapping season. (GMRT-D-35)

65. "Couldn't be better scene," said Don. (GMRT-D-36)

66. "Couldn't be better tree," said Don. (GMRT-D-36)

67. "Couldn't be better season," said Don. (GMRT-D-36)

68. More time than he could have saved would now be locked trying to get his bearings. (GMRT-D-40)

69. More time than he could have saved would now be sent trying to get his bearings. (GMRT-D-40)

70. Championship diving is the importance of such specifics as muscular control and coordination plus exact timing. (GMRT-D-42)

71. Championship diving is the spring of such specifics as muscular control and coordination plus exact timing. (GMRT-D-42)

72. Championship diving is the reading of such specifics as muscular control and coordination plus exact timing. (GMRT-D-42)

73. Championship diving is the result of such specifics as muscular springboard and coordination plus exact timing. (GMRT-D-43)

74. In 1959 the reverse side of the Lincoln cent was massed. (GMRT-D-45)

75. The wheat heads were published by a front view of the Lincoln Memorial, situated in Washington, D.C. (GMRT-D-46)

76. The wheat heads were registered by a front view of the Lincoln Memorial, situated in Washington, D.C. (GMRT-D-46)

77. The wheat heads were reversed by a front view of the Lincoln Memorial, situated in Washington D.C. (GMRT-D-46)

78. A windshield made of steel glass is relatively safe because the plastic layers have an elastic quality which prevents broken glass from shattering and causing injuries. (GMRT-D-48)

79. A windshield made of laminated glass is relatively safe because the plastic layers have an elastic quality which each broken glass from shattering and causing injuries. (GMRT-D-49)

80. A windshield made of laminated glass is relatively safe because the plastic layers have an elastic quality which tries broken glass from shattering and causing injuries. (GMRT-D-49)

81. A windshield made of laminated glass is relatively safe because the plastic layers have an elastic quality which encourages broken glass from shattering and causing injuries. (GMRT-D-49)

82. A windshield made of laminated glass is relatively safe because the plastic layers have an elastic quality which causes broken glass from shattering and causing injuries. (GMRT-D-49)

83. Language changes through the return of new words and the dropping of old ones. (GMRT-F-5)

84. These changes in language often plan changes in conditions within the community. (GMRT-F-6)

85. These changes in language often forego changes in conditions within the community. (GMRT-F-6)

86. Though a few minutes earlier I had felt that I could walk no further, the sight of the sparse landmark, the solitary tree, tonight silhouetted against the wintry sky, caused me to quicken my pace. (GMRT-F-7)

87. Though a few minutes earlier I had felt that I could walk no further, the sight of the familiar landmark, the solitary tree, tonight grouped against the wintry sky, caused me to quicken my pace. (GMRT-F-8)

88. By fixing an individuals place in society at birth, the caste system prevented many talented people from desirable positions where they could use their abilities for the benefit of the nation. (GMRT-F-16)

89. By fixing an individuals place in society at birth, the caste system prevented many talented people from successful positions where they could use their abilities for the benefit of the nation. (GMRT-F-16)

90. A foreign populated district in the North of Scotland is entitled to its programs as much as an industrial area. (GMRT-F-17)

91. A Scottish populated district in the North of Scotland is entitled to its programs as much as an industrial area. (GMRT-F-17)

92. A British populated district in the North of Scotland is entitled to its programs as much as an industrial area. (GMRT-F-17)

93. Immediately I knew her whom he spoke. (GMRT-F-24)

94. Oxygen can be prepared in the laboratory by combining potassium chlorate. (GMRT-F-29)

95. In such cases it is conceivable that the occurrence of large droplets into the base of the clouds or of artificial freezing bodies into the tops of the clouds might cause precipitation or at least hasten its occurrence. (GMRT-F-32)

96. In such cases it is conceivable that the elimination of large droplets into the base of the clouds or of artificial freezing bodies into the tops of the clouds might cause precipitation or at least hasten its occurrence. (GMRT-F-32)

97. In such cases it is conceivable that the cluster of large droplets into the base of the clouds or of artificial freezing bodies into the tops of the clouds might cause precipitation or at least hasten its occurrence. (GMRT-F-32)

98. This was most likely to occur in large, economically complex societies marked by unequal distribution of wealth and control by an active poverty. (GMRT-F-39)

99. For a man to be on good terms with himself and his neighbors, he must live in a society of equals where he depends not on the caprice of a strong and wealthy minority but on sovereigns applying to all members of the community establishing them. (GMRT-F-40)

100. For a man to be on good terms with himself and his neighbors, he must live in a society of equals where he depends not on the caprice of a strong and wealthy minority but on nations applying to all members of the community establishing them. (GMRT-F-40)

101. Some of Darwin's conclusions were so odd to accepted beliefs that they were condemned as absurd, contrary to common sense. (GMRT-F-41)

102. Some of Darwin's conclusions were so contrary to accepted beliefs that they were condemned as often, contrary to common sense. (GMRT-F-42)

103. Some of Darwin's conclusions were so contrary to accepted beliefs that they were condemned as completely, contrary to common sense. (GMRT-F-42)

104. Goods, objects with enjoyment of fulfillment are the natural fruition of the discofery and employment of means when the connection of ends with a sequential order is determined. (GMRT-F-51)

105. Goods, objects with thoughts of fulfillment are the natural fruition of the discovery and employment of means when the connection of ends with a sequential order is determined. (GMRT-F-51)

106. Goods, objects with uses of fulfillment are the natural fruition of the discovery and employment of means when the connection of ends with a sequential order is determined. (GMRT-F-51)

APPENDIX C

THREE SEMANTIC DIFFERENTIALS

SEMANTIC DIFFERENTIAL

I

Directions: Rate only the ideational character of the content, avoiding the influence of any other variables.

familiar	1	2	3	4	5	6	7	unfamiliar
little	1	2	3	4	5	6	7	much
intellectual	1	2	3	4	5	6	7	unintellectual
simple	1	2	3	4	5	6	7	complex
interesting	1	2	3	4	5	6	7	boring
profound	1	2	3	4	5	6	7	superficial
easy	1	2	3	4	5	6	7	hard
subtle	1	2	3	4	5	6	7	obvious
earnest	1	2	3	4	5	6	7	flippant
abstract	1	2	3	4	5	6	7	concrete
clear	1	2	3	4	5	6	7	hazy
strong	1	2	3	4	5	6	7	weak
personal	1	2	3	4	5	6	7	impersonal
masculine	1	2	3	4	5	6	7	feminine
emotional	1	2	3	4	5	6	7	unemotional
pleasant	1	2	3	4	5	6	7	unpleasant
serious	1	2	3	4	5	6	7	humorous
good	1	2	3	4	5	6	7	bad
precise	1	2	3	4	5	6	7	vague
informative	1	2	3	4	5	6	7	uniformative
formal	1	2	3	4	5	6	7	informal
general	1	2	3	4	5	6	7	technical

SEMANTIC DIFFERENTIAL

II

Directions: Rate only the language of the selection avoiding the influence of any other variables.

intellectual	1	2	3	4	5	6	7	unintellectual
easy	1	2	3	4	5	6	7	hard
subtle	1	2	3	4	5	6	7	obvious
succinct	1	2	3	4	5	6	7	wordy
earnest	1	2	3	4	5	6	7	flippant
clear	1	2	3	4	5	6	7	hazy
strong	1	2	3	4	5	6	7	weak
personal	1	2	3	4	5	6	7	impersonal
masculine	1	2	3	4	5	6	7	feminine
emotional	1	2	3	4	5	6	7	unemotional
pleasant	1	2	3	4	5	6	7	unpleasant
serious	1	2	3	4	5	6	7	humorous
florid	1	2	3	4	5	6	7	plain
good	1	2	3	4	5	6	7	bad
precise	1	2	3	4	5	6	7	vague
familiar	1	2	3	4	5	6	7	unfamiliar
little	1	2	3	4	5	6	7	much
simple	1	2	3	4	5	6	7	complex
interesting	1	2	3	4	5	6	7	boring
general	1	2	3	4	5	6	7	technical
formal	1	2	3	4	5	6	7	informal

SEMANTIC DIFFERENTIAL

III

Directions: Rate only the affective character of the content avoiding the influence of such variables as ideas and language.

thoughtful	1	2	3	4	5	6	7	thoughtless
simple	1	2	3	4	5	6	7	complex
profound	1	2	3	4	5	6	7	superficial
little	1	2	3	4	5	6	7	much
subtle	1	2	3	4	5	6	7	obvious
familiar	1	2	3	4	5	6	7	unfamiliar
clear	1	2	3	4	5	6	7	hazy
strong	1	2	3	4	5	6	7	weak
personal	1	2	3	4	5	6	7	impersonal
masculine	1	2	3	4	5	6	7	feminine
pleasant	1	2	3	4	5	6	7	unpleasant
serious	1	2	3	4	5	6	7	humorous
good	1	2	3	4	5	6	7	bad
precise	1	2	3	4	5	6	7	vague
affected	1	2	3	4	5	6	7	genuine

BIBLIOGRAPHY

BIBLIOGRAPHY

Allen, J., Jr. The Right to Read -- Target for the 70's. Paper presented at the Annual Convention of the National Association of State Boards of Education (California, 1969).

Alshan, C. M. A factor analytic study of items used in the measurement of some fundamental factors of reading comprehension. Unpublished doctoral dissertation, Teacher's College, Columbia University, 1964.

Anastasi, Anne. Psychological testing. New York: The Macmillan Company, 1961.

Anderson, C. C. A factorial analysis of reading. British Journal of Educational Psychology, 1949, 19, 220-221.

Auerbach, Irma-Theresa. Toward a model of reading comprehension. Unpublished qualifying paper, Harvard University, 1970.

Be A Better Reader. Teachers edition foundations A. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.

Bleismer, E. P. Reading abilities of bright and dull children of comparable mental ages. Journal of Educational Psychology, 1954, 45, 321-331.

Bormuth, J. R. Readability: a new approach. Reading Research Quarterly, 1966, I, 72-132.

Bormuth, J. R. New developments in readability research. Elementary English, 1967, 18, 830-845.

Bormuth, J. R. Empirical determination of the instructional reading level. In J. A. Figurel (Ed.), Reading and Realism. Newark, Delaware: International Reading Association, 1969a, p. 716-721.

Bormuth, J. R. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969b, 3, 358-367.

California Achievement Test (1963). See California Test Bureau 1957, 1967, and Tiegs, E. W. and Clark, W. W., 1963.

California Test Bureau. Technical Report on the California Achievement Tests. California: McGraw-Hill Book Company, 1957.

California Test Bureau. Technical Report on the California Achievement Tests. California: McGraw-Hill Book Company, 1967.

California Test Bureau. Catalog 1968-69. California: McGraw-Hill Book Company, 1968.

Chall, Jeanne S. This business of readability: a second look. Educational Research Bulletin, 1956, 35, 89-99.

Chall, Jeanne S. Interpretation of the results of standardized reading tests. In Helen M. Robinson (Ed.), Evaluation of Reading, Proceedings of the Annual Conference on Reading. Chicago, Ill.: University of Chicago Press, 1958a, pp. 133-138.

Chall, Jeanne S. Readability: an appraisal of research and application. Columbus, Ohio: Ohio State University, 1958b (Research Monograph, No. 34).

Chall, Jeanne S. Learning to read: the great debate. New York: McGraw-Hill Book Company, 1967.

Chandler, T.A. Reading disability and socioeconomic status. Journal of Reading, 1966, 10, 5-21.

Chihara, C.S. and Fodor, J.A. Wittgenstein: the Philosophical Investigations. Garden City, New York: Anchor Books, 1966.

Chomsky, Carol. The acquisition of syntax in children from 5 to 10. Cambridge, Mass.: The M.I.T. Press, 1969 (Research Monograph No. 57).

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press, 1965.

Cohen, Muriel. Fourth grade testing surprise. Boston Herald-Traveler, April 9, 1971, p. 1,5.

Coleman, J.S. et al. Equality of Educational Opportunity. O.E. 38001, Office of Education: U.S. Dept. of H.E.W., 1966.

Commins, W.D. and Fagin, B. Principals of educational psychology. New York: The Roland Press Company, 1954.

Cooper, Bernice. An analysis of reading achievement of white and negro pupils in certain public schools in Georgia. School Review, 1964, 72, 467-471.

Cronbach, L.J. Educational psychology. New York: Harcourt, Brace and Company, 1954.

Dale, E. and Chall, Jeanne. A formula for predicting readability. Educational Research Bulletin, 1948, 27 (1), 11-20, 28.

Davis, F. B. Fundamental factors of comprehension in reading. Psychometrika, 1944, 9, 185-197.

Davis, F. B. Research in comprehension in reading. Reading Research Quarterly, 1968, 3, 499-545.

Dykstra, R. The effect of code-and-meaning emphasis in beginning reading programs. The Reading Teacher, 1968, 22, 17-23.

Dyer H. See Stevens, W. K. (1971).

Farr, R. Reading: what can be measured? Newark, Delaware: International Reading Association, 1969.

Flesch, R. A readability formula in practice. Elementary English, 1948, 25, 19-26.

Forbes, F. W. and Cottle, W. C. A new method for determining readability of standardized tests. Journal of Applied Psychology, 1953, 37, 185-190.

Gans, Roma. A study of critical reading comprehension in the intermediate grades. Teachers College Contributions to Education, No. 811, 1940.

Gardner, E. F., Merwin, J. C., Collis, R., and Madden, R. SAT Manual High School Battery. New York: Harcourt, Brace and World, Inc., 1965.

Gates, A. I. Sex differences in reading ability. Elementary School Journal, 1961, 61, 431-434.

Gates-MacGinitie Reading Test. See Gates, A. I. and MacGinitie, W. H., 1965, 1969.

Gates, A. I. and MacGinitie, W. H. GMRT Teacher's Manual: Primary A. New York: Teachers College, Columbia University, 1965a.

Gates, A. I. and MacGinitie, W. H. GMRT Teacher's Manual: Survey D. New York: Teachers College, Columbia University, 1965b.

Gates, A. I. and MacGinitie, W. H. GMRT Teacher's Manual: Survey F. New York: Teachers College, Columbia University, 1965c.

Gates, A. I. and MacGinitie, W. H. GMRT Technical Manual. New York: Teachers College, Columbia University, 1965d.

Gates, A. I. and MacGinitie, W. H. GMRT Technical Supplement. New York: Teachers College, Columbia University, 1969.

Gates, A. I. and MacGinitie, W. H. GMRT Technical Supplement: Survey F. New York: Teachers College, Columbia University, 1970.

Glaser, R. and Cox, R. Criterion-referenced testing for the measurement of educational outcomes. In R. A. Weisgerber (Ed), Instructional process and media innovation. Chicago: Rand McNally, 1968, Pp. 545-550.

Guilford, J. P. New standards for test evaluation. Educational and Psychological Measurement, 1946, 6, 427-439.

Guttman, L. The structure of interrelations among intelligence tests. Proceedings of the 1964 Invitational Conference on Testing Problems, Princeton: Educational Testing Service, 1965, p. 25-36.

Guttman, L. and Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. Educational and Psychological Measurement, 1967, 27, 569-580.

Hall, W. E. and Robinson, E. P. An analytic approach to the study of reading skills. Journal of Educational Psychology, 1945, 36, 429-442.

Harcourt, Brace and World, 1968-1969 Catalog of Standardized Tests and Related Services. New York: Harcourt, Brace and World, 1968.

Harootumian, B. Intellectual abilities and reading achievement. Elementary School Journal, 1966, 66, 386-392.

Hilgard, E. R. Theories of learning. New York: Appleton-Century Crofts, Inc., 1956.

Holmes, J. A. and Singer,H. Speed and power of reading in high school. Cooperative Research Monograph No. 14, Office of Education, U. S. Department of H. E. W., 1966.

Horn, E. Methods of instruction in the social studies. New York: Charles Scribner's Sons, 1937.

Huey, E. B. (1908). The psychology and pedagogy of reading. Reissued. Cambridge, Mass.: The M.I.T. Press, 1968.

Hunt, L. C. Can we measure specific factors associated with reading comprehension? Journal of Educational Research, 1957, 51, 161-172.

James, W. On the function of cognition. Mind, 1928, 37, 1-18.

Jenkinson, Marion D. Basic elements of reading comprehension. Proceedings of the Second World Congress on Reading. Copenhagen, Denmark: (August) 1968, Pp. 41-47.

Johnson, M. S. Factors in reading comprehension. Educational Administration and Supervision, 1949, 35, 385-405.

Kelley, T. L., Madden, R., Gardner, E. F., and Rudman, H. C. SAT Directions for Administering Intermediate II Battery. New York: Harcourt, Brace and World, Inc., 1964a.

Kelley, T. L., Madden, R., Gardner, E. F., and Rudman, H. C. SAT Directions for Administering Intermediate I Battery. New York: Harcourt, Brace and World, Inc., 1964b.

Kelley, T. L., Madden, R., Gardner, E. F., and Rudman, H. C. SAT Directions for Administering Primary I Battery. New York: Harcourt, Brace and World, Inc., 1964c.

Kelley, T. L., Madden, R., Gardner, E. F., and Rudman, H. C. SAT Technical Supplement. New York: Harcourt, Brace and World, Inc., 1966.

Kendall, M. G. and Stuart, A. The advanced theory of statistics (Volume 2, 3). London: Charles Griffin and Co., 1966.

Kerfoot, J. F. Problems and research considerations in reading comprehension. The Reading Teacher, 1965, 18, 250-256.

Kingston, A. J. Reactions to theoretical models of reading: implications for teaching and research. In H. Singer and R. B. Ruddell (Eds.), Theoretical models and processes of reading. Newark, Delaware: International Reading Association, 1970, p. 183-186.

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

Klein, S. The uses and limitations of standardized tests in meeting the demands for accountability. Evaluation Comment, 1971, 2, 1-7.

Kolers, P. A. Introduction. In E. B. Huey, The psychology and pedagogy of reading. Cambridge, Mass.: The M.I.T. Press, 1968, p. xiii-xxxix.

Langsam, R. S. A factorial analysis of reading ability. Journal of Experimental Education, 1941, 10, 57-63.

Levin, H. and Williams, Joanna P. Basic studies in reading. New York: Basic Books, 1970.

Locke, J. (1697). Of the conduct of the understanding. In F. W. Garforth (Ed.), John Locke's Of the conduct of the understanding. New York: Teachers College Press, 1966.

Lorge, I. The Lorge and Flesch readability formulae, a correction. School and Society, 1948, 67, 141-142.

Marks, E. and Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Miller, G. A. The psychology of communication. Baltimore, Maryland: Penguin Books, Inc., 1967.

National Committee for Support of the Public Schools. The right to read, Ninth Annual Conference, 1971.

Neville, D., Pfost, P., and Dobbs, V. The relationship between anxiety and silent reading gain. American Educational Research Journal, 1967, 4, 45-51.

New England Right to Read Conference Prospectus, Andover, Mass., 1971.

New Practice Readers, A manual for the series, books A-G. New York: Webster Division, McGraw-Hill Book Company, 1962.

Otis, A. S. Otis Quick-Scoring Mental Ability Tests. Manual of Directions: Alpha. New York: Harcourt, Brace and World, Inc., 1954.

Pitcher, G. The philosophy of Wittgenstein. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964.

Reading Exercises, Manual of directions and answer key. New York: Teachers College, Columbia University, 1965.

Reading for Concepts. A guide for teachers: books A-H. New York: McGraw-Hill Book Company, 1970.

Richards, I. A. Practical criticism. New York: Harcourt, Brace and World, 1929.

Robinson, Helen M. The major aspects of reading. In Reading, seventy-five years of progress. Chicago: University of Chicago Press, 1966. (Supplementary Education Monograph, No. 96.)

Schlesinger, I.M. and Weiser, Z. A facet design for tests of reading comprehension. Reading Research Quarterly, 1970, 4, 566-578.

Simons, H. The relationship between aspects of linguistics performance and reading comprehension. Unpublished doctoral dissertation. Howard University, 1970.

Skinner, B. F. The distribution of associated words. Psychological Record, 1937, 1, 69-76.

Smart, B. H. Thought and Language. London: Longman, Brown, Green and Longmans, 1855.

Smith, F. Understanding reading -- a psycholinguistic analysis of reading and learning to read. New York: Holt, Rinehart and Winston, Inc., 1971.

Sochor, E. Elena. The nature of critical reading. Elementary English, 1959, 31, 47-58.

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

Standard Test Lessons in Reading, Teacher's Manual. New York: Bureau of Publications, Teachers College, Columbia University, 1950.

Stanford Achievement Test. See: Kelley, T. et al. (1964) and Gardner, E. F. et al., 1965.

Stanford Achievement Test, Directions for Administering. New York: Harcourt, Brace and World, 1964.

Stevens, W. K. Test experts call I.Q. and grade equivalency scores "monstrosities." The New York Times, March 23, 1971, p. 19.

Stewart, D. Esq. Philosophical Essays. Philadelphia, Pa.: Fry and Kammerer, Printers, 1811.

Strang, Ruth. The reading process and its ramifications. Paper presented as an invitational address to the International Reading Association: Newark, Delaware, 1965.

Sullivan, Elizabeth T., Clark, W. W., and Tiegs, E. W. California Short-Form Test of Mental Maturity. Examiner's Manual: Level 1. California: California Test Bureau, 1963a.

Sullivan, Elizabeth T., Clark, W. W., and Tiegs, E. W. California Short-Form Test of Mental Maturity. Examiner's Manual: Level 2. California: California Test Bureau, 1963b.

Taylor, W. L. Cloze procedure, a new tool for measuring readability. Journalism Quarterly, 1953, 30, 415-433.

Taylor, W. L. Cloze readability scores as indices of individual differences in comprehension and aptitude. Journal of Applied Psychology, 1957, 14, 19-26.

Thorndike, E. L. The measurement of ability in reading. Teachers College Record, 1914, 15, 207-277.

Thorndike, E. L. Improved scales for measuring ability in reading. Teachers College Record, 1915, 16, 445-467.

Thorndike, E. L. An improved scale for measuring ability in reading. Teachers College Record, 1916, 17, 40-67.

Thorndike, E. L. The psychology of thinking in the case of reading. Psychological Review, 1917a, 24, 220-234.

Thorndike, E. L. Reading as reasoning: a study of mistakes in paragraph reading. Journal of Educational Psychology, 1917b, 8, 323-332.

Thorndike, E. L. The understanding of sentences: a study of errors in reading. Elementary School Journal, 1918, 18, 98-114.

Thurston, L. L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Tiegs, E. W. and Clark, W. W. CAT Manual: Advanced. California: California Test Bureau/A Division of McGraw-Hill Book Company, 1963a.

Tiegs, E. W. and Clark, W. W. CAT Manual: Elementary. California: California Test Bureau/A Division of McGraw-Hill Book Company, 1963b.

Tiegs, E. W. and Clark, W. W. CAT Manual: Lower Primary. California: California Test Bureau/A Division of McGraw-Hill Book Company, 1963c.

Tremont, J. J. Cloze procedure as a method for measuring readability of text, reading comprehension of groups, and idea density in instructional materials. Unpublished qualifying paper. Harvard University, 1967.

Trenaman, J. M. Communication and Comprehension. London: Longmans, Green and Company, Ltd., 1967.

Vehar, Mary A. Extroversion, introversion, and reading ability. Reading Teacher, 1962, 21, 357-360.

Vernon, P. E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22, 269-286.

Wittgenstein, L. Preliminary studies for the philosophical investigations generally known as the Blue and Brown Books. Oxford: Blackwell, 1958.

VITA

1960-1964	The City College of The City University of New York	B.S., February 1964
1964-1966	Substitute Teacher, Common Branches New York City Public and Private Schools	
1964-1966	The City College of The City University of New York	M.S., June, 1966
1965-1966	Research Assistant, Education Clinic The City College of The City University of New York	
1966-1967	Research Assistant, Office of Research The City University of New York	
1967	Statistical Consultant Center for Urban Education, New York City	
1966-1968	The City College of The City University of New York	C.A.S., June, 1968
1967-1968	Lecturer, Office of Research The City University of New York	
1968	Project Director, Office of Research The City University of New York	
1968-1971	Grolier Fellow, Graduate School of Education Harvard University	