DOCUMENT RESUME

ED 054 233 24 TM 000 872

AUTHOR Bormuth, John R.
TITLE Development of Standards of Readability: Toward a Rational Criterion of Passage Performance. Final Report.
INSTITUTION Chicago Univ., Ill.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-9-0237
PUB DATE Jun 71
GRANT OEG-0-9-230237-4125(010)
NOTE 219p.

EDRS PRICE EDRS Price MF-$0.65 HC-$9.87
DESCRIPTORS *Cloze Procedure, *Criterion Referenced Tests, Decision Making, *Educational Accountability, Educational Objectives, Elementary Grades, High School Students, Instructional Materials, Models, Multiple Regression Analysis, *Performance Criteria, *Readability, Reading Comprehension, Scores, Standardized Tests, Test Construction, Testing, Testing Problems

ABSTRACT

The purpose of these studies was to develop and demonstrate a model for identifying criterion levels of performance that can be rationally defended as being the best level of performance for a particular instructional task. The specific objective was to identify the score on a cloze test that represents the most desirable level of performance on instructional materials. (Author)

Final Report

Project No. 9-0237
Grant No. OEG-0-9-230237-4125(010)

DEVELOPMENT OF STANDARDS OF READABILITY: TOWARD A RATIONAL CRITERION OF PASSAGE PERFORMANCE

John R. Bormuth

The University of Chicago

Chicago, Illinois

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.

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

Office of Education
Bureau of Research

ACKNOWLEDGEMENTS

Important contributions were made to these studies by a number of people. First recognition must be given to David E. Wiley, who not only advised on some of the problems in data handling but made several substantive contributions by clarifying some of the issues being dealt with. Important contributions were also made by graduate research assistants. Robert Bortnick managed much of the field work and data processing for Studies II and III. Patrick Finn's major contribution was in the area of item writing for the completion tests. Patricia Cocks' major contribution was to supervise the scoring of the completion tests. However, all three performed well the many other tasks that graduate assistants are supposed to perform. Martha Grossblat, the project assistant, made outstanding contributions to this project, helping to plan and supervise the testing in the field, organizing and supervising the clerical staff, editing and advising on the test materials and the final report, and generally executing with excellence all of the duties of a project assistant. Barbara Jill Buroker assisted her in much of this work, and Christopher S. Robbins managed much of the data processing and programming for Study III. Finally, I wish to acknowledge the wise counsel of George O'Speed on points of importance.

Special thanks are extended to the following school personnel for providing major assistance in carrying out these studies.

In the Hinsdale, Illinois, Public Schools:
Dr. Ronald Simcox, Superintendent
Dr. Lauren L. Schwisow, Assistant Superintendent
Buford Daniel, Principal, South High School
Charles LeCrone, Administrative Assistant, South High School
Florence Irwin, Head of English Department, South High School
Dr. Louis J. Adolphsen, Principal, Central High School
Clifford Derman, Principal (ret.), Junior High School
Ralph Larson, Principal, Junior High School
Thomas Burken, Assistant Principal, Junior High School
Irene Hanley, Girls' Counselor, Junior High School
Reno Bertollotti, Principal, Prospect School
Alfred Boysen, Principal, Madison School
Gertrude Matthias, Principal, Lane School
Raymond Shirk, Principal, Walker School

In the Downers Grove, Illinois, Public Schools:
Charles Elmlinger, Assistant Superintendent
Dr. Robert Neumann, Principal, South High School
Harold Mitchell, Guidance Director, South High School
Mary Ann Schorn, Guidance Counselor, South High School
Daniel Mahaffey, Principal, O'Neill Junior High School
Jacquie Jensen, Counselor, O'Neill Junior High School

Richard Bort, Principal, Hillcrest School
Rose M. Davis, Teacher, Hillcrest School
Norman Crandus, Principal, Indian Trail School
Vernon Langley, Principal, El Sierra School
Alexander Pawlowicz, Principal, Fairmont School

In the Glen Ellyn and Lombard, Illinois, Public Schools:
Florence Devine, Assistant Superintendent, District 44
Dr. Earl Dieken, Assistant Superintendent, District 41
James Verchota, Assistant Superintendent
John Sheahan, Prircipal, Glenbard West High School

J.R.B.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The purpose of these studies was to develop and demonstrate a model for identifying criterion levels of performance that can be rationally defended as being the best level of performance for a particular instructional task. The specific objective was to identify the score on a cloze test that represents the most desirable level of performance on instructional materials. Criterion scores have been used in the past to provide a rational means for deciding issues such as whether instructional materials are suitably easy for a student, whether the student has acquired enough knowledge on one unit of instruction to commence studying a more complex unit, and whether instruction should be stopped. However, the procedures for identifying these criterion scores have been, if not actually arbitrary, at least unexplicit and unrationalized. The thesis of the work reported here is that performance criterion scores can be identified by rational means.

The approach taken to this problem started with four fairly obvious propositions: (1) that instructional materials are studied because doing so yields positive benefits to the student and his society, (2) that this study is also accompanied by costs or negative benefits, (3) that these positive and negative outcomes of study vary as a function of the student's score on the criterion test over the instruction, and (4) that a reasoned approach to identifying the criterion level of performance would set the score at the performance level where a weighted sum of the outcomes showed that the maximum benefit was to be expected. The weights referred to in this summation should reflect how much each outcome from studying the instructional materials contributes to the ultimate outcomes of instruction and the relative values society places on the ultimate outcomes.

A fairly complete taxonomy of the variables affected by studying passages of instructional materials might include cognitive variables such as learning, retention, and transfer of information in the passage; proficiency variables such as rate of reading and latency of the responses acquired from the passage; affective variables such as the student's preferences for the subject matter, style, and difficulty of the passage and his willingness to study it; economic factors such as the costs involved in instructing the student to various levels of ability to cope with the passage and the costs involved in preparing materials suitable for him; and psychosocial factors such as the effects on the student's self-concept and attitudes from having him study materials at a particular level of difficulty relative to his own ability. In the studies reported here, the criterion selection model only included measures of information gain, rate of reading, willingness to study, and preferences for the subject matter, style, and level of difficulty.

The objectives of the data gathering aspects of these studies were (1) to obtain the regressions between each of these variables and cloze scores, (2) to obtain a set of weights representing the relative values placed upon each of these variables, and (3) to determine what variables influenced the shapes of the regressions and therefore required a differentiation of the passage performance criterion score. The studies were designed to permit the results to be generalized to students in grades 3 through 12, to materials on most of the topics and at most of the difficulty levels that these students are likely to encounter in instruction, and to each of the major purposes for which the students are likely to read a passage. Cloze and grade level consistently interacted in all of the regressions, thereby requiring that different criterion scores be identified for each grade level. Moreover, students assigned different ratings to materials depending on whether the materials were to be used for textbook, reference, or voluntary reading purposes, requiring that the criterion scores also be differentiated according to the use made of the materials. The weights assigned to each behavior were obtained by having teachers rate the relative values of each of the variables in the model. This was an expedient rationalized by the suppositions that teachers are acquainted with the goals of instruction, the values society places on those goals, and how the variables in the model contribute to the attainment of those goals. The teacher weightings did not differ between grade levels but did differ between uses of the material.

The regressions were entered into the criterion selection model, along with their weights and several statistical adjustments, to obtain the relative summed values of reading a passage at each of the levels of cloze performance. The point at which one of these curves peaked defined the passage performance criterion score for the grade level of student and use of the materials represented by the curve. Thirty criterion scores were obtained in this way, one for each of the three uses of materials at each of the ten grade levels of students. The criterion scores were then criticized for systematic biases and random error. The model in its present stage of development omits several important variables from the taxonomy of variables that should be represented, and could be said to be biased for this reason. Moreover, it appeared that the preference variables largely determined the points at which the criterion scores were set. This could be legitimate except that there was some evidence that the preference variables were highly redundant and, therefore, may have biased the criterion selection model. However, when the criterion scores were examined with respect to a measure of the efficiency rates of the variables in the model, the efficiency rates on most variables were at fairly acceptable levels at nearly all of the criterion scores. While much work remains to be done before fully acceptable criterion scores are obtained, the criterion scores obtained here appeared somewhat useful. But more important, it seemed that a fairly strong case had been made for the proposition that performance criterion scores can be identified by rational procedures that represent a consensus of how this type of decision ought to be made.

INTRODUCTION

Purpose

Some educators have greeted the concept of accountability with mixed feelings. On the one hand it seems clear that there should be an accounting from those entrusted with the education of the nation's children, with a very large role in determining the nation's future, and with the control of the massive amounts of public moneys spent on education. These matters are so important that they must be made the subject of public debate and public policy. On the other hand, because of earlier public neglect and indifference, the educators have had neither the need nor the means to develop the management techniques necessary to explicitly rationalize and account for educational decisions, instead relying on personal experience and subjective judgment. Thus, many feel that it would be difficult to defend many educational practices in open public debate, not necessarily because of any lack of merit in the practices themselves, but rather because the educator often lacks the management techniques necessary to arrive at those decisions in a rational manner that is open to public inspection and amenable to independent verification. Thus, though the accountability concept may have some advantages, efforts to incorporate it into practice could have some rather destructive effects.

One source of this problem is the fact that the development of evaluation systems has not kept pace with the changing role of education in American society. Up to the present time, achievement tests have been used primarily for ranking students relative to other students with whom they are in competition. To aid in this use of the test scores, much labor has been devoted to developing devices for referencing the test scores to the performances of norm groups: devices such as grade equivalent scores, percentile scores, age equivalent scores, and standard scores. These tests are primarily useful for selecting and rejecting students for admission to higher levels of education. Glaser (1963) has referred to tests of this type as norm reference tests.

However, as the nation underwent the social, economic, and technical changes of the last few decades, education took on new roles, roles that require major changes not only in the design of instruction but also in both the roles that tests play in that instruction and in the kinds of interpretations made of test performances. Tests intended for these new roles, called criterion reference tests, are designed for the additional purposes of (a) monitoring the student's learning during instruction so that the instruction can be altered as needed, (b) deciding when a student has gained sufficient mastery of the content he is

studying to warrant advancing him to a more complex unit of instruction, and (c) deciding when he has sufficiently mastered the content of instruction to enable him to perform adequately on the real-world tasks to which the instruction is relevant. However, at the present time there are few test interpretation devices to aid educators in making these judgments from scores on criterion reference tests.

In broad terms, the objectives of the present studies are to develop the concept of a rationally derived performance criterion for use in interpreting scores on criterion reference tests and then to explore the practical and theoretical problems encountered in actually trying to identify such a criterion. But in more specific terms, the criterion attempted is that level of cloze test performance that can be rationally defended as representing the optimal, the most desirable, or the literate level of performance on passages drawn from instructional materials. I shall refer to this criterion as a passage performance criterion. While no claim will be made that this performance criterion was identified with finality, a fairly useful one was identified. But, possibly more important, a considerable understanding was achieved of both the nature of the problem of identifying performance criteria rationally and the kinds of procedures that must be developed.

Purpose of a Passage Performance Criterion

Scaling Readability: The original motive for developing a passage performance criterion arose out of the research in readability assessment. The basic objective of readability research is to develop a theory of reading comprehension by analyzing the linguistic features that operate as stimuli for the comprehension processes. And one of the applied objectives of this work is to develop regression formulas that permit educators to estimate how much reading comprehension skill a student must have in order to exhibit an acceptable level of performance on a given set of instructional materials. These estimates are obtained by identifying certain linguistic features of a passage such as the number of relative pronouns or the average sentence length, inserting these variables for the unknowns in the regression equation, and then solving to obtain the readability estimate for the passage. This readability estimate is useful for selecting appropriate materials provided that it tells the educator how much ability a student should have in order to comprehend that passage at an acceptable level, and provided that this estimated ability is expressed in terms of ability measures readily available to the educator.

At the time the studies reported here began, highly accurate formulas had just become available, but those formulas had little practical value because they express passage readability in a metric that is meaningless to an educator, that is, in terms of the predicted cloze difficulties of the passages. Tests made by the cloze readability procedure (Taylor, 1953) had been introduced into readability research because

they provided a number of important advantages over earlier methods of measuring passage readability. Previously, investigators used tests composed of comprehension questions and found the mean percentage of items students answered correctly. However, tests of this type are subject to unpredictable variations in the size of the mean scores, variations that are due to the uncontrolled ways test writers select and phrase the items included in the tests. As a result investigators could never be certain what they were measuring--varitations in the readabilities of their passages or merely variations in the behaviors of test writers.

Tests made by the cloze readability procedure, the specific type of cloze procedure employed throughout these studies, were free of these effects. They are made by replacing every fifth word in a passage with an underlined blank of a standard length. Students who have not read the intact passage are asked to write in each blank the word they think was deleted, and their responses are scored correct only when they exactly match the word deleted. The research on tests of this type has now grown too extensive to review here (see Rankin, 1965; Bormuth, 1967b; and Potter, 1968, for extensive reviews); however, it shows that the processes underlying cloze test responses are seemingly indistinguishable from those underlying ordinary comprehension tests and that the cloze readability procedure is a highly efficient method of making and scoring the tests.

Major improvements in the accuracy of readability formulas were made when cloze tests along with modern psycholinguistic theory were applied to the study of readability (Bormuth, 1966 and 1969a, and Coleman, 1968). The formulas derived exhibited validity coefficients of .93 and higher. However, as mentioned earlier, these formulas produced passage readability estimates in a metric that was essentially useless to an educator. The formulas were calculated by regressing the linguistic measurements of passages on their cloze means. Consequently, the educators who applied these formulas would obtain readability estimates expressed in terms of the mean cloze score likely to be obtained by administering cloze tests made from that passage to the subjects in the populations the investigators used. Obviously, this is of little assistance in determining whether the passage is suitable for a particular student.

However, a method has been devised for scaling passages in terms of the reading achievement levels on standardized tests (Bormuth, 1969a). The procedure consists in administering both a cloze test and a reading achievement test to the students, regressing their reading achievement scores on their cloze test scores, and then using this regression equation to calculate the reading achievement scores of students who were able to score exactly at some performance criterion on the cloze test. This grade placement readability score for a passage can be interpreted by an educator as representing the amount of reading ability a student must have in order to comprehend the passage tested at the criterion

level of performance. Moreover, these readability grade scores can be incorporated into the calculations of the formulas so that passage readabilities are expressed in grade scores. However, the value of this or any other scaling procedure depends upon the validity of the cloze performance level accepted as the criterion. And, unfortunately, none had been established for cloze tests. The sequence of studies reported here set out to establish a performance criterion for use with cloze readability tests.

Assessment of Comprehension Literacy: A passage performance criterion has also turned out to be an essential component in a procedure that can provide a rational means of assessing a student's comprehension literacy (Bormuth, 1970a). Although the major justifications for free compulsory education have generally been to produce an electorate that can understand printed language well enough to inform itself on public issues and to perform essential economic tasks, we have never developed a means of assessing when a person has reached this level of reading comprehension skill.

Four types of literacy criteria have been widely used (Allen, 1969). The Bureau of the Census classifies a person as functionally literate if he has completed at least five years of school. Not only does this criterion ignore the fact that many people who have achieved this criterion cannot display more than the most rudimentary reading skill, but it also ignores the fact that no one knows whether even the best students after five years of school attendance can perform at an acceptable level on reading tasks that are essential to adults. A second criterion often employed in the public press is that a student should obtain test scores at the national norms for his age group on standardized tests of reading achievement. Since these norms represent the mean or average score for students of his age, the criterion is absurd since exactly half of all students would thereby always be classified as illiterate regardless of how well or how poorly they read. A third type of criterion often used is that the person be able to read as well as the average student at some age or grade, grade 6 for example. This does not represent much improvement over the other criteria since it still is not known at what grade level, if ever, the average student can read well enough to comprehend adult reading materials at an acceptable level. The fourth type of criterion classifies a student as literate if his reading achievement is equal to that of the average student who both is in the same grade and has the same degree of aptitude or intelligence. Not only does this criterion make an insidious and possibly incorrect assumption that the average child at each level of aptitude has learned to read as well as he possibly can, but it also ignores the central issue--how well a person must be able to read in order to cope with essential adult reading tasks.

A means has been proposed (Bormuth, 1970a) for rationally defining adult comprehension literacy using only techniques that are currently available. However, an essential component of this procedure is a

criterion of passage performance that is, itself, rationally established; that is, a criterion that can be shown to represent the most desirable level of comprehension performance on a passage. The main features of this literacy assessment procedure will be described in the final section of this report.

Rationale of Performance Criterion Scores

The justification for criterion reference tests and for the performance criterion scores used to interpret scores on those tests rests on three points. First, the place of education in American society has changed over the past several decades from that of a luxury to be enjoyed by a select few to that of a necessity for any individual in order to function in this society. Second, this changed role of education has resulted in the need to make basic changes in the design of instruction, from designs that induced much variability in learning and permitted only the very able elite to survive to the higher educational levels, to designs that produce a high level of subject-matter mastery in every student possible. Third, these new models of instruction require, in turn, changes in the functions achievement tests are designed to perform, from norm reference tests designed primarily to rank a student with respect to the other students with whom he is competing so that the most able can be selected for further instruction, to criterion referenced tests designed primarily to supply the information necessary for managing instruction, so that as many students as possible will achieve sufficient mastery of the subject matter to perform competently on relevant real-life tasks.

Early Context of Instruction: Historically, norm referenced tests have served to identify elite groups who are singled out to receive special recognition and rewards such as praise and high grades for their achievement and special encouragement and financial aid to proceed to higher levels of education. At the time the technology of norm referenced testing was being developed, there was considerable justification for performing these functions in American educational institutions: most industries had not yet developed advanced technologies and did not depend on a large and well-educated work force; social institutions and services were fewer and relatively accessible to an individual regardless of his educational attainment; and the means of production were relatively inefficient, producing little more than the necessities of life for most people. Whatever surplus wealth was produced had to be devoted mainly to increasing capital resources, and little was left to spend on what were considered non-necessities such as education. Thus, since at that time it seemed neither necessary nor economically possible to educate a large proportion of the population to high levels, it could be argued that norm referenced tests performed a valuable service by helping to reduce the numbers of people seeking advanced education and helping to select just the students who could be efficiently educated.

Changed Context of Instruction: However, just as the society itself has been changing rapidly over the past several decades, so have the roles education plays within society: much of our industry is now based on fairly advanced technologies and service professions, requiring a highly trained work force of a size that comprises a substantial proportion of our population; our social institutions have become so numerous and complex that a poorly educated person has great difficulty gaining access to their benefits or even exercising his franchise effectively; and, finally, not only do our means of production depend on a well-educated work force, but they are sufficiently efficient to produce the huge surpluses of wealth necessary to finance that education. Thus, education has ceased to function in the society as an institution of secondary economic and social necessity and has become, instead, an integral part of the mechanisms by which the society achieves its social and economic goals.

New Designs of Instruction: These changes in the functions performed by education have led to increasingly pressing needs to alter the basic design of instruction. Some of the changes required seem to have been treated most comprehensively by Carroll (1963) and Bloom (1971). They point out that instruction should now be designed to take every person possible to a high level of mastery of the instructional content. They would permit variability to occur among students, not in their level of achievement of the content, but only in the amount of time necessary to attain the level of mastery prescribed. This is not to say that they are proposing that everyone should be educated through the level of the Ph.D. or even that we should eliminate selectivity of the students who are to proceed to successively higher levels of education. Rather, these seem to be expressions of the view that a large amount of knowledge must now be regarded as essential for every member of the society who is to be a productive and constructive member of the society, and that instruction should be designed to produce in students nothing short of the mastery of this content. This new conception of the function of instruction has found expression in many of the current activities in educational research and development: programmed instruction and individually prescribed instruction, to name two. But for the purposes of this discussion, only the ramifications for evaluation will be dealt with.

New Designs of Tests: The proposed changes in the design of instruction, in turn, require that achievement tests perform new functions. Glaser (1963) argued that tests designed merely as norm reference tests were inappropriate for helping to make the decisions necessary in instruction that is designed along the new lines being considered. He pointed out that traditional tests were designed primarily to discriminate among students and rank them within their cohort groups, but that this information was irrelevant to many of the decisions that are now required. Those are the decisions of whether the student had achieved sufficient mastery of the content on one instructional unit to justify elevating him to the next more complex unit, whether the instruction itself was sufficiently effective or needed revision, and whether the student had achieved a sufficient level of mastery of educational content to permit him to function

at a desirable level on relevant real-world tasks. Again, this is not to say that anyone has proposed doing away with norm reference tests, but rather that the newer conceptions of instruction require the use of achievement tests designed to perform additional functions, functions that traditional norm referenced tests are ill-designed to perform.

Traditional Performance Criteria: The concept of a performance criterion is one of the major components of new instruction designs that has yet to be submitted to rational and empirical analysis. This should not be taken to mean that no one has ever employed performance criteria in making instructional decisions. Quite the contrary, numerous examples can be cited: compulsory school attendance laws often set the ages of 6 and 16 as the minimum criteria for initiating and terminating instruction, if longevity can be said to be a performance; instructional programmers often continue to revise and improve their program until 90 percent of their students can answer 90 percent of the items on the program's terminal tests; many reading teachers avoid using a book in a student's instruction unless he can answer at least 75 percent of the questions they ask him about a sample passage taken from that book; the passages used to calculate readability formulas are often given grade equivalent difficulty values by finding the grade level at which the average student answers at least 75 percent of the questions on a test over that passage. However, it seems fair to characterize all of these performance criteria as being arbitrarily chosen. That is, their proponents did not offer reasoned arguments and evidence that performance at the criterion was any more desirable than any of the other performance levels that they could have selected as their criterion.

Need for Rational Performance Criteria

The arguments supporting the new models of instruction depend critically upon an assumption that has so far seemed to elude the writers on this subject. This is the assumption that performance criteria must be established through rational procedures, as opposed to the unrationalized or arbitrary procedures used to select the performance criteria presently employed. This is not to say that the present criteria were not based on a reasoning process of some sort, but rather that that reasoning process, if any, was not made formally explicit, supported by evidence, and open to critical examination and challenge. The dependence of the new models on this assumption is critical in the sense that the fundamental objectives of those models cannot be achieved unless the performance criteria are arrived at by rational means.

The argument supporting this contention begins by pointing out that two axioms are basic to the new deisgns: first, that education is now a primary means by which society achieves its goals, and second, that the design of instruction should reflect this fact. It follows then that the decisions made during the design and management of instruction should reflect society's resources, its goals, and the relative values it places

on each. Since performance criteria are employed in making these decisions, the criteria must be arrived at by rational means, provided that setting a given performance criterion at various levels affects the degrees to which society's goals are reached and its resources are consumed. If it can be shown that varying the levels of performance criteria has these effects, then it becomes evident that performance criteria employed in the new designs of instruction must be arrived at by methods that rationalize the criteria in terms of those goals and resources, and the priorities set on each.

The data obtained in the present studies provide direct evidence that variations in the level at which a performance criterion is set does influence the costs and benefits of instruction. However, a more general support is required for this argument. Since there are at least three distinctly different categories of performance criteria and since the present studies relate only to a special case of one of these categories, this argument must deal separately with each of the three major classes of performance criteria. For convenience in discussing them, these three types of performance criteria will be referred to by the terms instruction termination criteria, unit termination criteria, and response level criteria.

Instruction Termination Criteria: An instruction termination criterion is the performance level on a test that is used to determine whether a student's formal instruction in a subject matter area should be terminated. For example, this might be a particular score on a test of reading comprehension. Students reaching or exceeding that score would receive no further instruction in the reading comprehension skills, while students falling below it would receive further instruction. It seems fairly clear that the level at which such a criterion is set would have a marked effect upon the level of costs society would incur from instruction. If, for example, the criterion were set at a zero performance level and if that level of performance literally meant that a person had no measurable amount of knowledge or skill in the subject matter mentioned, then the costs of instruction would be very low. Similarly, if the criterion were set at perfect performance on the test and if that performance meant that the student could correctly exhibit every measurable skill or item of knowledge in the subject being considered, then the cost of instruction in that content area might be very great, especially if that content area were a major one such as reading or mathematics.

Virtually every learning curve ever published has shown that, although learning occurs very rapidly on the first few exposures to the material, the rate diminishes rapidly on each successive repetition and levels out well before perfect performance is reached. Moreover, there is a large constant component in the cost of each repetition: the costs associated with teacher time, the use of capital and expendable equipment, and the like. As a result, setting criteria at perfect performance seems likely to make the instruction a very costly proposition. In any case, even this common-sense analysis provides fairly persuasive support for the proposition that varying the level of an instruction termination criterion also varies the level at which society's resources will be consumed.

Varying the level of an instruction termination criterion also appears likely to affect the benefits produced by the instruction. Society provides the means of instruction because it anticipates that by doing so it will derive some benefit, such as the building of an informed and responsible citizenry and the building of a work force capable of efficiently manning the technical aspects of industry. If a zero criterion were adopted in content areas such as the social studies, the sciences, mathematics, or reading, it goes without saying that society would also receive relatively few of the benefits that could be anticipated to result from that instruction. Conversely, if a very high performance criterion were adopted, society could expect to receive more of the benefits associated with the instruction. Although we have less understanding of how instruction influences affective learning, it seems reasonable to expect variations in the performance criterion to have effects on important affective as well as cognitive learning, affective learning such as enjoyment gained from applying the content learned. And so affective learning must also be considered in setting a performance criterion.

Just how much society would benefit from a high criterion level of performance undoubtedly depends also upon what kinds of tasks the instruction prepares a person to perform and how much demand there is for people to perform those tasks. But, barring the special case in which there were no demand for performance of the tasks that instruction prepares a person to perform, it seems fairly certain that varying an instruction termination criterion affects the level of benefits society can expect to obtain from instruction. Consequently, in setting the level of instruction termination criteria, we must rationalize the procedure employed, presumably by developing a model that would select the level at which the values received from instruction were maximized while the costs were simultaneously minimized. Unless we did so the instruction would not necessarily conform to society's goals and might actually require the use of society's resources in ways that prevented it from reaching its goals.

Unit Termination Criteria: The unit termination criterion is a particular score on a test over some subdivision or unit of the content normally taught in some subject matter area. The function of this score is to decide when a student's instruction in the unit should be terminated. It would be a very difficult matter to attempt to rationalize performance on a unit termination test directly in terms of the benefits to society of the skill tested. The skills measured by such tests often have, at best, a remote relationship to the performance of real-world tasks. For example, it is not immediately apparent how one would place a value of this kind on performance on a test of ability to pronounce one-syllable words ending in a final e or on a test of ability to subtract pairs of numbers in which the student is not required to "borrow." Rather than attempt to rationalize a unit termination criterion in these terms, it appears more reasonable to rationalize it in terms of its costs or benefits in attaining the instruction termination criterion.

It seems fairly certain that varying the level at which a unit criterion is set does have some rather large effects of this kind. Block

(1970) furnished direct evidence when he instructed several groups on a sequence of units in matrix algebra, holding each group to a different performance criterion. The group held to a 95 percent criterion exhibited higher scores on an achievement test given at the termination of the sequence and exhibited greater observed transfer in terms of the amount of time saved in studying each successive unit. However, those scores were only slightly higher than those exhibited by a group held to an 85 percent criterion. Moreover, the group held to the 85 percent criterion voiced a greater enjoyment of the learning experience and a greater desire to pursue the study of matrix algebra in the future than either the 95 percent group or the groups held to criterion scores of less than 85 percent. We could add to Block's observation the fairly reasonable common-sense argument offered by Bloom (1971) that a person held to very low criteria on unit termination tests might never reach a fairly high instruction termination criterion, especially if the content area were a highly structured one such as reading, mathematics, or a foreign language. The point, then, seems fairly well sustained that variations in the level at which a unit termination criterion is set probably affect the costs and benefits associated with attaining the relevant instruction termination criterion, and thereby it can be said that the level at which a unit criterion is set affects the way society utilizes its resources and the degree to which it attains its goals.

Response Level Criteria: A response level criterion is a particular score based on tests of the student's responses to the materials used during the instruction itself. This criterion is used to determine whether the instruction is at an appropriate level of difficulty for the student. For example, it is fairly common for instructional programmers to revise the segments of instruction on which the response rate falls below a criterion level. Similarly, teachers use readability formulas and informal reading tests to select materials on which the student will exhibit a response rate that they deem appropriate. Indeed, the passage performance criterion scores sought in the present studies are examples of the response level criterion. It appears difficult to rationalize response level criteria directly in terms of society's priorities for its goals and resources, the skills and knowledge measured being too remotely connected to those goals and resources. Rather, these criteria seem more properly rationalized in terms of their effects on the attainment of the behaviors sought at the termination of units and of instruction.

The present studies provide evidence that variations in response levels on cloze tests over instructional materials are associated with rather large effects on measures of information gain, rate of reading, and interest in the materials being studied, variables which determine the costs and benefits associated with attaining the behaviors sought at the termination of instruction. Thus, there seems to be both theoretical and experimental justification for the claim that variations in the level of a response level criterion determine the costs and perhaps even the feasibility of attaining the unit termination criterion, the instruction termination criterion, and consequently the goals of society and the way resources must be used.

We have now examined what seem to be the three major functional categories of performance criteria. Although direct evidence is scarce, it was possible to develop for each type of performance criterion a fairly plausible argument to the effect that variations in the level at which a criterion is set either directly or indirectly influence the degree to which society's goals are attained and the way its resources would be consumed in attaining those goals. Therefore, it seems justified to say that instruction cannot reflect the fact that it is one of society's principal means of attaining its goals unless the educational decisions based on performance criteria can be shown to conform to society's goals, its resources, and the priorities it places on each. This goal is attainable only if performance criteria are, themselves, justified in terms of their effects on those goals and resources.

Utility of a Model

It seems essential to select a performance criterion through the use of a model, a procedure that makes explicit the reasoning used and the way the evidence was combined to arrive at the criterion. The argument supporting this proposition leads from two points made in the preceding discussion: first, that deriving a performance criterion is an exercise in social policy-making, and second, that deriving a performance criterion cannot be considered as merely a technical problem since the ways society attains its goals and utilizes its resources are affected by the level at which the criterion is set. What is being brought to issue here is the question of whether the logic and evidence involved in setting a criterion should be set forth in the form of an explicit model. This question is approachable as a problem in ethics or in political philosophy, or merely as a pragmatic problem, any one of which would provide sufficient grounds for employing a rational model for identifying performance criteria.

Several arguments are possible to support this proposition from the point of view of ethics. To briefly develop just one, consider a broadly accepted principle that is applied in medical practice and that many would like to require of the practices of any institution dealing with the public, such as advertising and manufacturing companies. This principle holds that neither society nor the individual should be subjected to an action unless they first understood and accepted its possible effects--consequences for health or for damage to the environment, for example. Since the level at which a criterion is set has consequences that are undoubtedly unforeseen by noneducators, this ethical principle is violated unless the considerations relevant to the derivation of a criterion are made known to the public. When these considerations are made completely explicit and formalized, they, in fact, represent a model.

Arguments having the same result can be developed from the point of view of political philosophy. In a democracy, for example, it is held that the power to make social policy should reside ultimately in

the hands of its citizens. And social institutions such as schools are developed to carry out those social policies. However, these policies can be either accidentally or intentionally subverted if the managers of these institutions are allowed to make decisions which affect the demands on society's resources or the degree to which its goals are attainable. Consequently, the managers of such institutions must be held accountable for all such decisions. Since the level at which a performance criterion is set affects society's goals and resources, it follows that the managers of educational institutions must make as clear as possible the basis upon which they set their performance criteria so that the matter can become a conscious policy. Failure to try to develop and present such a model for inspection and policy consideration represents, in effect, a violation of a major tenet of democratic political philosophy.

Perhaps the most telling argument is the purely pragmatic observation that the public will probably never provide more than marginal support for an institution that is not clearly productive and responsive to its demands. A common complaint among educators, for example, is the reluctance of many voters to support bond issues and tax increases for education, even to the point where some schools have been forced to close temporarily for want of funds. When this fact is coupled with the observation that the principal support for the notion of educational accountability has not come from professional educators, we obtain some grounds for believing that some of the public's parsimony arises from an uncertainty about what and how much it is getting for its money. The use of performance criteria based on explicit rationales would help to inform the public of what it is getting in return for its support and to inform them of the uses to which proposed increases in expenditures would be put.

CRITERION SELECTION MODEL

Possibly the most useful result of the studies being reported here was the conception of a performance criterion as something that one could identify rationally with a model. The work on this model has generally proceeded at a much more rapid pace than the data collection operations, and so the model that can be employed with the data available is somewhat less complete than the best model that could be devised at this time. Consequently, there is a need to present separately the model actually employed with the data and the subsequent conceptual improvements in the model. This will be done by first developing the model actually used and then presenting suggestions for its improvement in the form of a critique of that model. This section will give only minimal attention to the instruments used to operationalize the model.

Function of the Criterion

This passage performance criterion is designed to function as a response level criterion for use with written instructional materials. When it has been developed to a satisfactory state, educators will be able to use it to determine whether materials are suitable for use in a student's instruction. They may do this either directly or indirectly. In order to make direct use of the criterion, they can make a cloze test from a sample of the materials they are using in instruction and obtain the student's score on that test. If his score approximates the passage criterion score, it may be interpreted as representing the most desirable level of performance, and the materials would thereupon be judged as suitable for use in his instruction. If his score fell above or below the criterion score, it would be interpreted as showing that he was receiving less than the maximum benefit from those materials; that is, that some important values were being sacrificed by using those materials for his instruction--whereupon his instruction would be altered by altering the readability of the materials, by selecting more appropriate materials, or by providing him with a different amount of preparation before he studied the materials. One indirect use consists in first using the criterion to determine how much ability a student must display on the norms of a norm reference test in order to read each of a set of passages, developing readability formulas that predict the ability levels these passages require, and then applying these formulas to determine the suitability of materials for a given student. Another indirect use would incorporate this criterion into a procedure for assessing comprehension literacy that would establish a termination criterion for literacy instruction.

Logic of the Model

The object of the model, then, was to formalize the reasoning process by which one would identify the most desirable level of response

on cloze tests made from written, verbal instructional materials. This reasoning process included six major points.

Outcomes of Reading: In order to determine the cloze test performance level that yields the most desirable returns, it is obviously necessary to decide what types of responses occur during reading. These fall roughly into three categories--cognitive, proficiency, and affective or interest. The cognitive response type selected for this model was the information the student gained while reading. This is measured using short-answer completion and multiple-choice questions to test a student both before and after he has read a passage and then using the difference between these two scores. The reason for using a gain score instead of just the score on a comprehension test given after the student has read a passage is that some of the variance in the latter test scores cannot be attributed to the reading of the materials. People often have some prior knowledge of the area discussed by a passage. The test items themselves add to this information because the question stems in many cases repeat fragments of the passage. This knowledge enables students to answer some questions directly and some by inference. Consequently, in order to determine what value a student has gained from reading a passage, it is necessary to subtract from his score on a post-reading test his score on a pre-reading test, correcting the difference for biasing effects, of course.

The measure of proficiency used was the reading rate, measured in words per minute, that the student exhibited in reading a passage. Rate of reading is valued for two reasons. First, it represents a measure of the amount of labor required of the student in order to gain the information he was able to get. Second, it represents a measure of the efficiency with which his time was used. In either case, rate can be regarded as a benefit or at least as a negative cost of reading a passage.

Students' affective responses were measured using a set of seven-point preference rating scales. The student was asked to rate a passage for how well he liked to learn about the subject matter the passage contained, how well he liked the writing style used, how suitable the level of difficulty was for him, and how willing he would be to study those materials. It seems important to include affective measures of this sort for both short- and long-range considerations. In the short range, affect is important because it may index the likelihood that a student will actually attend to the materials. Obviously materials that students do not like and try to avoid studying are apt to be less than effective. A long-range consideration is the possible effects that disliked materials have on a student's willingness to continue study in a field and perhaps on his willingness to stay in school or to go into occupations that depend upon his knowledge in that field. These expectations gain some credibility from the argument that such materials constitute an aversive stimulus that is made to occur repeatedly in conjunction with the content being studied. Thus, it would be plausible to expect that adverse responses to the materials would in turn be conditioned to occur in response to

the area of knowledge contained in those materials, causing the student to avoid the further study and use of that area of knowledge.

Selection of Base Variable: It is possible to identify a performance criterion for any one of the outcomes associated with reading a passage; all that is needed is an appropriate model. Consequently, it is necessary to select some one of these variables in which to express the criterion and then to defend that choice. It seems that this choice has to be made in terms of the primary function of the materials, selecting the variable that represents the primary behavior the materials are designed to produce. Thus, if the materials were designed primarily to influence students' attitudes toward various aspects of some subject matter, as in propaganda and moral education materials, then one would logically measure that as the primary outcome and ask for a model that answered the question, how much of these attitudes should the student acquire from the materials? If the materials were designed primarily to increase the student's proficiency in skills already acquired, as in review materials, then one would logically measure that as the primary outcome and ask for a model that answered the question, how much proficiency should a student acquire from the materials? For the present studies, it was assumed that the primary purpose of most instructional materials is to transmit information and skills to students, and therefore the decision was made to express the performance criterion in those terms and to develop an appropriate model for selecting the performance criterion on that measure.

Considered from just this narrow theoretical point of view, this reasoning would lead to the selection of a measure of information gain as the variable on which to identify the performance criterion. However, consideration of the difficulty and expense this would involve for a practical educator would lead to a rejection of information gain as a base variable. In order to measure information gain it is first necessary to make a comprehension test for a passage, second to administer it both before and after the student has read the passage, and third to calculate a difference score that is corrected for the biases inherent in gain scores, that is to calculate a residual gain score. It seems that the best test to make is a comprehension test composed of verbal questions of the familiar types. However, producing a good quality of question requires skills that are normally acquired only through considerable training and experience, and writing the questions is a very time-consuming job. In addition it will be shown below that in order to produce tests that reflect just the properties of the passages and not the idiosyncrasies of the test writer, even more highly technical skills are required. Also, the administration of the test must be carried out following designs that prevent the first administration of the test from influencing the student's performance on the second administration, designs that require both technical skill to develop and much labor to execute. Finally, the calculation of a residual gain score requires some knowledge of statistics and rather massive amounts of calculation. Thus, taken individually or together, these considerations seem sufficient to show that expressing

a performance criterion in terms of a gain score would result in an evaluation system whose use would demand resources quite beyond those available in schools and impose costs that would make its use a highly questionable matter.

For these reasons it seemed advisable to replace information gain with a surrogate variable that was both easy to measure and highly correlated with gain. Information gain was then considered one of the outcomes associated with this surrogate base variable. Since gain would be highly correlated with the surrogate, gain would automatically be weighted statistically in the model as if it were the base variable. Tests made by the cloze procedure were selected for this role. They are easily and inexpensively made and they are highly related to measures of comprehension and information gain.

Dependence on Cloze Scores: The model proposes to identify the performance criterion on cloze tests by taking advantage of the fact that various levels of cloze performance are associated with different levels of performance on each of the desired outcomes of reading a passage. The logic here is fairly direct: The value of reading a passage is judged in terms of what a student is expected to get from the passage, that is in terms of the expected outcomes. And the value of exhibiting a particular level of cloze score while reading the passage is determined by the expected levels of each of those outcomes associated with this cloze score relative to the levels of those outcomes corresponding to other cloze scores. For example, students who obtain a cloze score of 10 percent on a passage normally obtain a very low level of performance on a test of information gained from that passage, while a cloze score of 60 percent is normally associated with a much higher level of information gain. The model asserts that, with respect to information gain, a cloze score of 60 percent should be assigned a value that is proportionately higher than the value assigned to a cloze score of 10 percent. The model requires that a set of such values be assigned to each cloze score, one value for each outcome (such as information gain) from reading the passage.

To state this matter in operational terms, the values of cloze scores, with respect to an outcome, can be expressed by the curve that represents the regression of that outcome on the cloze scores. However, these values do not represent purely relative values of the cloze scores since the regression curve may not pass through the origin and, possibly, might even take on negative values. However, they can be converted to relative values if the cloze scores are standardized. The model deals only with values that are relativized in this manner.

The use of regression equations in this manner has an important weighting effect on one outcome relative to the weight given other outcomes that is desirable but might go unnoticed unless it is pointed out explicitly. The logic of this model does not claim that all of the variance of a given outcome, interest for example, is attributable to the base variable, cloze scores. Quite the contrary, the model specifically

asserts that, in identifying a performance criterion, an outcome should be given only a weight relative to the other outcomes that is proportionate to the degree to which the variance in that outcome may be attributed to variations in cloze performance. This weighting is automatically reflected in the amount of variation in a regression line and requires only a minor adjustment to compensate for unreliability in the tests. It should be explicitly recognized, however, that a weighting operation of this sort is a part of the model.

Teacher Weightings: Presumably, each of these types of outcomes has a different value for the student. Their relative values were determined by asking teachers to rate the relative values of these behaviors for coping with the reading assignments they give in their instruction. Then each behavior was assigned a weight that was proportional to its mean rating.

Combination of Response Levels: Finally, it was possible to determine for a given cloze score the level of information gained, the rate of reading, and the levels on each of the preference rating scales. And the level of each of these behaviors could be weighted by multiplying it by the mean weight teachers gave that behavior. When this was done, the sum of these weighted levels provided an estimate of the general value obtained by reading a passage at that level of cloze performance. When the same operations were performed for all cloze scores, it was possible to select the cloze score having the highest general value. This level of cloze performance was taken as the response rate criterion for passages.

Differentiated Criterion Scores: In the course of these studies it became evident that a single response rate criterion might not be appropriate for all students or for all of the uses to which instructional materials are put. Consequently, separate criterion scores were calculated for students at different grade levels. And, within each of those grade levels, a separate criterion was calculated for each of the three major functions for which materials are used--textbook reading, reference reading, and voluntary reading. The only implication this had for the tests described above was for the rating scales. The student was asked to rate a passage as part of a textbook, as a reference, and finally as a book that he might read voluntarily.

Formalization of the Model

In reducing this model to a formal statement, the first step was to define percentage scores on cloze tests as the independent variable. Thus, the performance criterion score was to be expressed in terms of the cloze metric. Scores on tests of all of the other response types were then defined as dependent variables, each of these dependencies being expressed by means of a multiple regression equation that included higher powers of cloze scores in order to fit the curved relationships

that were common in the regressions of these variables. The student's grade level was also used as an independent variable in many of these regressions. However, it will be omitted here for the sake of simplicity in the immediate presentation. The regression equations, then, were of the following form:

$$(1) \qquad R = \alpha_0 + \alpha_1 C_i + \alpha_2 C_i^2 + \cdots + \alpha_x C_i^n \quad .$$

In equation (1) the dependent variable is the rate score, R, the independent variable is the cloze score, C, and the constants of the parameters are represented by α, with α_0 representing the intercept and α_1 to α_x the parameters associated with various powers of the score, i, on the cloze tests. If the relationship between rate and cloze were linear, only the first two terms of the equation would be needed, of course. However, if that regression were curved in some way, the equation would contain whatever powers of the cloze score were necessary to fit the curve to a satisfactory degree of approximation. For convenience in this discussion, these equations will be represented in the abbreviated form

$$(2) \qquad R = f(C_i) \quad ,$$

where $f(C_i)$ stands for the right-hand side of equation (1).

Equations of the type shown in (2) can be used to represent the regressions of the scores on cloze tests and the scores on each of the other response types tested. This would obtain expressions (3) through (8):

(3) Information Gain, $I\text{-}G = f(C)$

(4) Rate, $R = f(C)$

(5) Subject Matter Preference, $S\text{-}M = f(C)$

(6) Style Preference, $S = f(C)$

(7) Difficulty Preference, $D = f(C)$

(8) Willingness-To-Study, $W\text{-}T\text{-}S = f(C)$

Each of these equations can be used to determine the level of any of these valued response types for any given cloze score.

The model eventually calls for us to find the total of these values for each cloze score. This can be expressed in this fashion:

$$(9) \qquad V_i = f(C_i)_{I\text{-}G} + f(C_i)_R + f(C_i)_{S\text{-}M} + f(C_i)_S + f(C_i)_D + f(C_i)_{W\text{-}T\text{-}S} \quad .$$

The notations in this equation are read, the general value (V_i) associated with a cloze score of size C_i is the function relating information gain to that cloze score plus the function relating rate to that cloze score, and so on.

However, before we can perform such an addition we must assign appropriate weights to each of the response types. There are at least two major aspects to these weighting operations. The simplest of these is the weight assigned to remove arbitrary effects introduced when the test scores are expressed in terms of their original metrics on the tests used to measure them. For example, rate is expressed in its raw metric as the number of words read per minute; those scores can vary over a wide range, in these studies from roughly 60 to 400 words per minute. The performance ratings, on the other hand, were obtained using seven-point scales. If these scores were summed in the model using their raw score metrics, rate of reading, for example, would receive a weight of perhaps 5 to 1 relative to each of the response types measured on the preference scales. The arbitrary nature of this effect may be seen from the fact that this weight ratio would have been reversed had the intervals on the performance scales been assigned the numbers 1000 to 7000. The solution was to express the scores as deviation scores. These deviation scores are represented by the use of lower-case letters as shown in (10).

$$(10) \qquad r = f(C_i) = f(C_i)_r$$

If the summing operation were performed after the variables had been standardized, however, we would obtain an equally unacceptable result, for the variables now would be given exactly equal weights. This would be unacceptable because the variables are obviously not of equal value for attaining the behaviors measured on a unit termination test. Consequently, their relative values for this purpose must be determined and corresponding weights assigned. In this case the weights were based on the teachers' ratings of the relative values of each behavior.

A number of other adjustments must also be made to the data to allow for biases arising from regression effects and the effects peculiar to the designs of the specific studies. These adjustments will be discussed in relation to the analyses of specific sets of data.

The model in its final form, but disregarding these statistical and design corrections, is as follows:

$$(11) \qquad V_i = [W_{i-g} \cdot f(C_i)_{i-g}] + [W_r \cdot f(C_i)_r] + [W_{s-m} \cdot f(C_i)_{s-m}] + [W_s \cdot f(C_i)_s] + [W_d \cdot f(C_i)_d] + [W_{w-t-s} \cdot f(C_i)_{w-t-s}] \; .$$

This expression is read, the general value of reading a passage at a given cloze level, i, is equal to the function relating cloze scores to information gain at cloze score i, $f(C_i)_{i-g}$, multiplied by the weight given information gain, W_{i-g}, plus, and so on.

Evaluation of the Model

This model remains incomplete in several respects. Its current deficiencies are owing, in part, to the fact that when the author began this work, he thought he was working on a rather narrow and well-defined problem in the scaling of passages for readability analyses. Only after the major data collection operations had gotten under way did the full generality of the problem as it is presently stated become apparent. However, in even larger measure, the deficiencies in the present model arise from fundamental inadequacies in our conceptions of what behaviors ought to be produced by instruction and in our knowledge of how those behaviors are related to each other. These problems must be solved en route to the development of the models for selecting performance criteria. The chief motive for publishing the model at this time was that it had become apparent that the job was too large for a single individual to perform. Indeed, this paper will have served its major purpose if it defines the problem well enough to enlist the interest of others in helping to seek its solutions.

This critique deals mainly with problems arising out of our lack of systematic theories of what behaviors instruction can produce and of how those behaviors relate to each other. A number of interesting methodological problems were also encountered in the design of the tests and in the analyses of the data. However, these matters are largely specific to the particular operations chosen, and so they will be discussed in those contexts.

Taxonomy of Behaviors: A model of this sort is fully defensible only if it includes measures of all of the important behaviors the instruction produces. Omission of any variable can potentially affect the point at which V_i maximizes and thereby bias the performance criterion. The present model does not include every important behavior and, possibly, no model can at the present time, for we do not have a systematic theory of what behaviors are produced by instruction. However, the model could have been considerably more complete than it is. Considering the cognitive response types first, measures of transfer and long-term retention clearly should have been included. One would ordinarily expect a considerable degree of transfer from one section of a book to subsequent sections, since the concepts dealt with in one section are often developed in an earlier section. Moreover, as a student learns the author's methods of organization and styles of presentation in the earlier sections, this knowledge should aid him in dealing with subsequent sections. Presumably, students would ordinarily differ in the degree to which they acquired this knowledge from a passage and in the degree that they were able to apply it in subsequent sections to obtain the transfer

effects, that is, to study more rapidly and to perform better on tests than students who had not studied the earlier passage. The reasons for including measures of long-term retention seem self-evident.

Measures of proficiency--that is, response latency, response availability, or speed of a correct response--have been largely neglected in instructional theory and nearly neglected in this model. Clinical observation has led the author to a strong suspicion that the ease with which a student is able to call up the knowledge he has acquired has a strong influence on how much he appears to have retained on recall tests, on how well he is able to apply this knowledge in tests and actual situations where he must use that knowledge for application and evaluation purposes, and finally on the degree of transfer of that knowledge to new learning situations. While the rate of reading variable actually included in this model is probably related to measures of response availability, the relationship is indirect and ambiguous. It seems likely that, for example, a person can "complete" a passage either rapidly or slowly while having acquired much of its content to either a low or high level of latency. A more direct measure might be obtained by measuring the duration of time between the presentation of a question about the passage and the onset of a correct response. Presumably, variability in such latency measures would be related to cloze performances on passages and would consequently be relevant to establishing a criterion performance level on them.

The preceding omissions could have been avoided simply by employing classical learning theory as a device to taxonomize response types. In that body of theory a response is said to have four major attributes: It is acquired, forgotten, and transferred, and it has a latency. Another major class of omissions occurred when information gain was treated as being a single variable. Educators (Bloom, 1956) claim that the responses themselves can be differentiated into homogeneous classes such as inference, application, and evaluation, depending on how they had to be acquired. That is, some of the knowledge acquired from a passage is obtained through processes that cannot be explained as responses merely to the explicit structural features of the language in the materials. However, the test of information gain used in this model included only tests of the information explicitly signaled.

It is relatively more difficult to evaluate the completeness of the affective tests represented in the model, because there has been less study of the role of affect in instruction. One valuable addition to the present model was suggested by Block (1970), who argued that if the instruction were so difficult that the student reached only a low level of mastery on it, this would cause him to regard himself as less competent to deal with the type of content being studied and that this, in turn, would cause him a certain amount of anxiety about pursuing study in that area. Block found support for this argument when he observed that students whose instruction was adjusted to aid them in reaching a unit termination criterion of 85 percent on units of matrix algebra

expressed a greater amount of interest in pursuing this area of content than students who were trained to lower performance criteria. But even with the addition of this affective variable, the taxonomy of affective tests seems to be the most uncertain aspect of the taxonomy of variables in the model.

The next problem was that of distinguishing in a reasoned way variables that are both relevant to performance on a passage and sufficiently important to include in a model from variables that are also relevant, but not important enough to include in a model. For example, in reading instructional materials in content areas such as history of science, information gain and rate seem both relevant and important to achieving the objectives of the instruction. However, as the student reads, he also employs skills such as sounding out words, page turning, looking up the meanings of words, and the like. These behaviors are relevant in the sense that achievement of the instructional objectives would be either partially or wholly blocked if the student could not perform them. But the behaviors may or may not be important to the attainment of the criterion, depending chiefly upon whether or not the student has mastered them. Thus, a variable is important to the attainment of a criterion only if it exhibits some variability in the students with whom the criterion is to be used, and then it is included only if it is relevant to the independent variable, that is if it exhibits a correlation with the independent variable, the metric in which the criterion score is expressed.

In establishing a passage criterion, then, a behavior such as ability to pronounce, or call, the words in the passage would probably be considered important for students up to about grade 6 but possibly of no importance for students in higher grades. It could also be conjectured that such skills as sounding out a word interfere with the comprehension of a passage whenever they must be employed. Thus, a measure of the latency of word-calling might exhibit importance for students at every level. On the other hand, a measure of page-turning ability seems unlikely to exhibit importance for students at any but the very earliest levels of instruction.

Weighting Variables for Importance: The most difficult problem to deal with in the model seems to be the problem of assigning weights to each of the variables. There are two aspects to this weighting problem. The first is the easiest to dispose of. That aspect arises out of the consideration that not all of the variance in a given variable can be attributed to variations in performance levels on the cloze tests, and so the variable would be given an inappropriate weight if this were not taken into account. This was done in the model. The dependent variable scores were first transformed into true scores to adjust their variances for differences in variance due to differences in test reliability. Then the regression equations were fitted. The degrees of variability in the different regression curves were then approximately proportional to the proportion of each variable's variance that could be attributed to the variance of cloze scores.

However, a practical problem that will be more difficult to solve occurs when we consider how important a behavior might be. Asking teachers to rate the importance of the dependent variables is an expedient that merely begs the questions of why each of those behaviors is important and what criteria will have to be used to establish their relative importance. It seems clear that the behaviors achieve their importance from the fact that they can ultimately be rationalized in terms of their contribution to the attainment of the performance criteria we establish for the termination of instruction. Consequently, these weights may ultimately have to be based on studies that determine what these contributions are.

Summation of Variables: This model requires us to add together the weighted values of all of the dependent variables as if they were totally unrelated to each other. This, in fact, was not the case. For example, it would appear plausible to argue that how a person would rate his willingness to study a passage would probably be dependent to some degree on how interested he was in its content and how well he could understand it. Evidence will be presented to show that substantial correlations exist among these variables. Disregarding these covariations has the effect of adding some variances into the model two or more times, giving them a spurious weight in determining the performance criterion. In order to allow for these effects, it is first necessary to develop the theory of hierarchic relationships among these variables.

MEASUREMENT PROBLEMS

Instrumenting this model raised anew some old but basic theoretical problems that have received almost no analysis in the context of criterion reference measurement. Consequently it is necessary to analyze them as a preface to the present studies. This discussion will describe precedents for the use of performance criteria in education and analyze some of the basic psychometric assumptions that must be met by the tests used in the present studies. There are a number of fairly close analogies between norm reference testing and the form of criterion reference testing used here. These analogies were described throughout these discussions in an effort to clarify the contrasts that are of central interest.

Precedents for Performance Criteria

In some respects, establishing a passage performance criterion must be regarded as an unorthodox procedure from the standpoint of classical psychometric theory. Glaser (1963) pointed out that the main thrust of the activities in classical psychometric theory has been to identify the attributes or response types that are useful for discriminating among students and to develop test designs for making these discriminations efficiently. In the present studies, however, the object is to identify the types of responses that are useful for making discriminations among passages and then to develop test designs for making these discriminations efficiently with respect to as few as a single student. Moreover, the conventional objective of measurement has been to compare people with other people, presumably with the assumption that the higher a person's score is, the greater will be his benefits. In the present studies the object is to compare his performance against a performance criterion, following the assumption that even good things like high test scores become undesirable if they cost too much. But in spite of these and a number of other contrasts that can be drawn, the work reported here is not entirely without precedent.

Cut-Off Criteria: It has long been a common practice in such areas as college admissions and job placement to establish and use performance criteria. In studies on college admissions it is not uncommon to encounter complex regression formulas that predict college grades, along with contingency tables that show the probability of college failure as a function of measures of student aptitude. Often, the statistical techniques are sufficiently rationalized to be regarded as at least a partial model of the process of accepting and rejecting applicants. The cut-off points derived in these studies must in any case be regarded as performance criteria. This is not to say that they are strictly analogous to the performance criteria developed in the present studies. For example, the conventional procedure attempts to predict a single dependent variable such as college failure from one or more independent variables, while in the present studies the effort is to locate the optimum level of one

outcome, the independent variable, by assessing its value in terms of a number of dependent variables. Moreover, the cut-off points themselves seem to have always been arbitrarily selected, arbitrarily in the sense that the particular score at which the cut-off was set was identified by some consideration external to the model. However, the basic functions of the two kinds of cut-off scores are at least analogous--to discriminate and select appropriate instruction as compared with discriminating and selecting applicants.

Passage Performance Criteria: Nor is the concept of a passage performance criterion new. Perhaps one of the earliest instances of criterion reference testing is a procedure that has long enjoyed wide use in reading instruction, the informal reading inventory. This procedure makes use of passage performance criteria on tests of both word recognition and comprehension. In most versions of this procedure, a student is asked to read a passage that is thought to be representative of a book or other materials, and then to answer some questions about the passage. If he is able to answer at least 75 percent of the questions, the materials are said to be at his instructional level and suitable for use in his supervised instruction. If he is able to answer at least 90 percent of the questions, the materials are said to be at his independent level and suitable for use in his unsupervised study and voluntary reading. Also, elaborate procedures are employed for evaluating the competency of younger students to cope with the word recognition demands of the materials.

The procedures used in the informal reading inventory are recommended in many of the major teacher training textbooks (see Betts, 1946; Bond and Tinker, 1967; and Harris, 1961, for examples), and there are teacher training courses in some universities devoted primarily to training teachers to use them. Moreover, the passage performance criteria used in these procedures were also employed routinely in the early readability formulas (Lorge, 1939; Dale and Chall, 1948; and Flesch, 1943) for scaling passages in terms of a grade placement metric.

Previous Studies: The precursors of the present attempt to derive a rational passage criterion started with the objective of finding out what scores on cloze tests were comparable to the 75 and 90 percent criteria used with ordinary comprehension tests. In the first study (Bormuth, 1967a) the students were given cloze tests over each of nine passages and later given multiple-choice comprehension tests over the same passages. Regression procedures were then used to determine that 45 and 52 percent on cloze tests were comparable to the 75 and 90 percent criteria, respectively. In a later attempt to replicate these results using similar procedures, Rankin and Culhane (1969) obtained the cloze values 41 and 61 percent. In a second kind of approach, students were asked to read passages orally and then to respond to short-answer completion questions. The cloze criterion scores obtained in this study were 44 and 57 percent, respectively. Subsequent analyses of the data from the two earlier studies show that some of the variability among the results

of these studies may have been due to a failure to take ceiling effects into account in calculating the regressions.

The present studies were undertaken when a search of the literature failed to reveal evidence that anyone had ever tried to validate the traditional criterion scores of 75 and 90 percent. Moreover, a broader search showed that there did not even seem to be an instance of anyone having developed a rational model for establishing performance criteria for use in managing a student's instruction.

Psychometric Assumptions

The tests used to establish or assess a passage performance criterion seemingly must meet at least three assumptions that are not ordinarily the source of much concern in classical psychometric theory. However, these assumptions demand attention here because the objectives of this type of criterion reference testing, in an important sense, reverse the objectives in norm reference testing. In norm reference testing the objective is to discriminate among students with respect to the extent to which they have learned from a single body of instruction. To measure these distinctions a set of items is made from the instruction to form a test that fairly represents the responses that could potentially be acquired by the various students. And then this test is given to several students in order to observe the distinctions of interest--the individual differences among students. Thus, in norm reference testing the instruction and its test are regarded as fixed while the students must be varied in order to observe the distinctions of interest. Conversely, in this form of criterion reference testing the objective is to discriminate among bodies of instruction with respect to the extent to which their respective contents are learnable for a single student. To measure these distinctions a set of items is made from various bodies of instruction to form a set of tests that fairly represents the responses that could potentially be acquired from the various bodies of instruction. And then these several tests may be given to as few as a single student in order to observe the distinctions of interest--the differences among passages. Thus, bodies of instruction and their respective tests are variable. For convenience in these discussions the three assumptions that are critical because of these contrasts will be referred to as the instructional conformity, the behavioral consistency, and the regression identity assumptions.

Instructional Conformity: The instructional conformity assumption asserts that the test made from a passage must represent just the characteristics of the passage itself, and not some other and irrelevant considerations. In operational terms this is the assertion that a student's score on the test should be affected only by the characteristics of the passage itself and not by any other source of systematic variance. Lorge (1949) identified the test writer as a major and uncontrolled source of variance that often works to defeat the objectives of this kind of testing. Accepted practices among test writers are to omit writing items

that in their judgment would be so easy or hard that they would not aid in discriminating among students and then, after trying out the tests, to cull out any remaining items of the same type. Moreover, if more items are needed to form a sufficiently reliable test, it is regarded as permissible to alter the phrasing of easy or difficult items to bring them into the desired range. Whatever may be the merits of these operations for making norm reference tests, and those merits are distinctly dubious (Bormuth, 1970b), their effect in criterion reference testing is to subvert the objective of discriminating among passages: on a difficult passage most of the items would be discarded or altered because they were too hard, thus making the passage appear easier than it is; on an easy passage most of the items would be discarded or altered because they were too easy, thus making the passage appear harder than it is; and the total effect would be to systematically bias the tests against exhibiting the variance of interest, the between passage variance. Applying these procedures to criterion reference tests, then, is roughly analogous to selecting for norm reference tests just the hard items that have a negative correlation with total test score and just the easy items that have a positive correlation with total test score. The effect would be to shrink and confound the variance of interest, the between subject variance, in that form of measurement.

A problem of a different order arises from the practice of omitting items merely because, in the opinion of the test writer, they measure knowledge that has little worth. This practice is inadmissible in any form of testing on both ethical and rational grounds, because the test writer is thereby arrogating the prerogatives of the subject matter and curriculum experts without having to account specifically for his acts, and without having to show the credentials that allow him to misrepresent the instruction. The practice is also inadmissible on purely psychometric grounds. The object of this form of criterion reference testing is to measure the characteristics of a passage, and the information contained in that passage is one of its characteristics. So the test is invalid to the extent that it arbitrarily excludes some of that content from being tested, regardless of what might be the social utility of the knowledge that is measured as a result. This is not to say, however, that certain classes of passage characteristics cannot be excluded from testing, for it is inconceivable that we would ever want to test every type of response that could possibly be made to a passage. Rather, it is the reasonable assertion that when a test is described as measuring a certain class of responses to a passage or containing a certain class of items, it should do just that and not include an insidious selection procedure that further selects within that class of responses in some possibly systematic but unaccountable manner. If nothing else, this issue can be viewed as a matter of honesty in labeling.

The tests used in the present studies met the instructional conformity assumption as well as the current technology of testing would permit. The cloze tests presented no problem since the cloze procedure itself defines the population of possible items, and the procedure for sampling

from this population does not permit the test writer to inject his judgments. In one sense the interest scales also met this assumption quite well. When each scale is considered a separate test, as was the case in these studies, it seems likely that each measured impartially whatever features in a passage could affect that scale. If, on the other hand, the entire set of scales is regarded as a single instrument that tested some behavior such as interest in the passage, per se, then there would undoubtedly have been biases in the selection of the scales because the scales actually used probably do not test everything, or even an unbiased sample of everything, in a passage that affects its interest. However, since each scale was treated as a separate test, this bias is not hidden among the unlabeled items of a test but is open to inspection and, indeed, was discussed as a bias in the tests selected for the model. These studies also included the use of comprehension tests made from questions of the short-answer completion and multiple-choice types. The procedures devised to insure their instructional conformity were so elaborate that they will be described only in conjunction with the descriptions of the specific studies. However, it seems likely that these tests did exhibit a fair conformity to the passages.

Behavioral Consistency: Whereas the concept of instructional conformity referred to the relationship between a passage and its test, the concept of behavioral consistency refers to the relationships among the behaviors measured by tests that are of the same type but that are made from different passages. This assumption asserts that tests bearing the same label but made from different passages must measure the same type of behavior on all of the passages. This assumption is required by the definition of a passage performance criterion. The definition regards passages as being comparable along various dimensions; that is, the passages are similar in that they can all elicit the same classes of behaviors and are variable in these respects only in the number or strengths of the responses they can elicit. This assumption is exactly analogous to the assumption that makes normative comparisons among students possible: Students are regarded as similar in that they can all make the same classes of responses and variable only in the number or strengths of the responses they can make.

In order to at least nominally meet the behavioral consistency assumption, essentially the same controls are exercised in both criterion and norm reference testing. The kind of response elicited by an item is thought to be primarily determined by its relationship to the instruction. That is, the kind of response elicited by a single question can be varied by holding the question constant and varying the instruction it tests. As an example, suppose that a list of words were presented to two students for them to pronounce. If one student had never before seen these words but had been trained in phonics skills, his responses would normally be regarded as an index of his phonics skill. If the other student, on the other hand, had no training in phonics but had been exposed to the words during look-and-say instruction, the identical set of items would ordinarily be regarded as measuring his sight recognition

skills. The reason for regarding the test differently for these two people, then, is that the relationship between the items and the instruction of the individuals differed. When criticism is leveled at the validity of the standardized achievement tests widely used in American education, it is often because of some sort of failure to meet the behavioral consistency criterion for all the students tested, as when teaching for the tests changes the items intended to test problem solving into rote responses or when curriculum innovations in some school district cause items that previously tested only simple problem solving to require complex reasoning processes. In norm reference testing, the method of controlling behavioral consistency is by giving all students the same items or at least items sampled from a common population and then by cautioning test users against making comparisons among students who have followed materially different instructional programs.

In the form of criterion reference testing used here, this particular method of control is impossible since each passage represents a different piece of instruction and since the tests must also differ. However, the necessary control can be exercised by making the same type of items for each passage. Bloom (1956) was among the first to explore in some detail the theory that the same classes of behavior can be tested on a wide range of instructional materials by controlling the type of item used. Bormuth (1970b) has further developed this theory, holding that items can be regarded as being of the same type only if they can be derived from instruction by the same set of operations, and he advanced the claim that items derived by the same set of operations would exhibit behavioral consistency between instructional programs. The research reported by Hively, Patterson, and Page (1968) and Bormuth, Manning, Carr, and Pearson (1970) tends to support his claim.

There is evidence that the particular tests used in these studies meet the behavioral consistency criterion. Factor analyses of tests made by the cloze procedure (Weaver and Kingston, 1963, and Bormuth, 1969b) have shown that cloze tests made from a variety of passages yield analogous results with respect to whatever factors emerge in the studies. The work on operationally defined comprehension items is still in the preliminary stages; however Bormuth et al (1970) defined a large number of classes of comprehension items, made one item from each of a number of different paragraphs, and then demonstrated that items derived by the same rules exhibited behavioral homogeneity in spite of the fact that they were written from different passages. When Singer (1969) reviewed the literature on comprehension tests made by traditional techniques, he found that the tests measured the same factors consistently even though a wide variety of materials had been used in these studies. Finally, there seems to have been only one previous study (Carroll, 1960) of the properties of rating scales when they are applied to passages. Carroll's factor analyses showed that his rating scales consistently measured the same behaviors across passages. Finally, Osgood, Suci, and Tannenbaum (1957), who developed the semantic differential as a means for measuring some aspects of the meanings of words, showed that

rating scales made by their procedures consistently exhibit the same factor loading patterns regardless of the particular samples of words rated on those scales. In addition, the present studies will present fairly extensive evidence that the tests used in them met the behavioral consistency assumption.

Regression Identity: The regression identity assumption asserts that a given type of test used to establish a performance criterion must exhibit parametrically identical regressions on the independent variable, the cloze tests, regardless of the passage from which it is made or the student to whom it is given. To illustrate, consider that a cloze and a rate test are given for each of several passages, that the rate scores are regressed on the cloze scores for each passage separately, and that the regression curves are all plotted on the same graph. This assumption requires that, within reasonable margins for error, these regression lines must coincide. This assumption is implicit in the model that defines the passage performance criterion. The logic of the model states that reading a passage produces multivariate outcomes that are correlated with cognitive performance measured by cloze scores, that those outcomes can be given weights corresponding to their relative importance, that the value obtained from performing at a given cloze score level is the weighted sum of these outcomes at that cloze score, and that the performance criterion is the cloze score at which these sums reach a maximum. Thus, if the regressions for some type of test differed for each passage, the criterion scores would also differ and there would be no such thing as a passage criterion per se, but rather there would be a different criterion for each passage. Similarly, if the regressions differed in the parameters for groups, a different criterion would have to be used with each group. As it turned out, the parameters of the regressions did differ for students at different grade levels in these studies, and so different performance criteria were necessary for each grade level of student.

The regression identity assumption is made in a somewhat more limited form in some applications of norm reference testing, where it is sometimes assumed that the regression between tests observed on one group of students is identical in parameters to the regression that would be observed in a similar group. For example, students' grades in college are sometimes regressed on their scores on the entrance exams they took earlier. This regression is then used to identify the level of ability on the entrance examination at which student failure rates reach unacceptable proportions, and this performance level is thereafter used as a cut-off in admitting and rejecting students. This application of the regression curve necessarily assumes that the same regressions calculated on successive groups of students are identical in their parameters. The chief distinction between the assumption as it is made in this example and as it is phrased for the present studies is that in the present studies the identity must also hold across all tests of the same type made from different passages.

Prior to the studies reported here and their unreported pilot studies, there was only fragmentary evidence that the regression identity assumption could be met. In the earliest study of this sequence (Bormuth, 1967a), the scores of nine multiple-choice tests were regressed on the scores from cloze tests over the same passages. The nine individual regression lines throughout the entire ranges of scores all lay within 4.2 percentage score points, on the comprehension tests, of the single regression line fitted to the entire set of tests. It seems possible that the regression lines would have coincided even more closely had operational item-writing procedures been used to insure that the behavioral consistency criterion had been met more closely. The data from the study using the orally read paragraphs (Bormuth, 1968) were also examined, though somewhat less formally. Each of a student's completion test scores was plotted against the cloze score he made on other paragraphs at that level of difficulty, thus providing a curve for each student. The curves of all 120 students were plotted on the same graph, resulting in a figure that gave the visual impression of a fairly sharply defined normal ogive curve with the individual student curves nearly all lying in a fairly compact band.

Coleman (1968) and Kamman (1966) provided evidence of a somewhat less direct nature. Coleman obtained a gain score obtained by two administrations of a modified form of cloze test from each of a number of passages. He then plotted this mean gain score against the mean cloze score for each passage. The result was roughly an inverted U-shaped curve. Kamman, on the other hand, measured various kinds of preferences for poems and the cloze difficulties of the poems. When he plotted mean preferences from each preference scale against the ranked cloze difficulties of the poems, his curves were generally of the inverted U-shape. However, the Kamman and Coleman data are difficult to interpret with respect to the issue at hand. Since they were based on group means, the underlying regressions for each passage could have been of any of several forms, sharp inverted V-shapes for example, that averaged out as inverted U-shaped curves merely because mean individual differences of students were ignored.

Since these studies depended critically upon the regression identity assumption being met and since there was little direct evidence on the matter, a number of increasingly elaborate pilot studies were conducted, the final one closely resembling the main studies in design. These pilot studies will be reported here in some detail. Among other things, they showed that the regression identity assumption seemingly can be met by tests of the type finally selected to instrument the model.

STUDY I

Purpose

The purposes of this first study were to determine how closely, if at all, the regressions between the information gain and cloze measures obtained from different passages resembled each other, and also to examine the shapes of those regressions. These purposes stand in contrast with the more common type of psychometric study, where the object is often to identify the mental processes underlying the responses to tests. In that kind of study, the correlation is used to measure the degree of overlap between tests. The objective of the present studies, on the other hand, is to make policy decisions based in part on the relative amounts of information gain associated with each level of cloze performance; and these decisions depend not only on the correlation between the variables but also on estimates of the parameters of their regressions.

At the time this study was conducted, it was uncertain whether a performance criterion could be established in the manner proposed. In order to do so, it is essential that the tests exhibit regressions that are identical or at least highly similar in their parameters, regardless of the passages from which they are made. A search of the literature revealed only the tangential evidence already discussed. Moreover, it seemed logically necessary that the tests exhibit both a high degree of conformity to the instruction and behavioral consistency as prior conditions for exhibiting regression identity. Meeting these assumptions seemed to require, in turn, that tests over different passages be made in highly similar ways. The cloze tests could meet these assumptions fairly easily. However, the procedures for constructing comprehension test questions relied heavily upon the judgments of a test writer. In this sense, then, this study served as an initial test of the feasibility of identifying a passage performance criterion.

Previous experience with cloze and multiple-choice tests made from the same passages (Bormuth, 1962) had suggested that these regressions would take on a form similar in shape to a normal ogive curve. The cloze tests appear to be both more difficult and at the same time capable of discriminating over a broader range of student abilities than multiple-choice tests. On an easy passage, multiple-choice score distributions often show marked ceiling effects resulting from many students obtaining perfect and near-perfect scores, and on difficult passages the distributions exhibit floor effects resulting from some students obtaining zero and near-zero scores. While cloze scores also exhibit a floor effect, they seem to do so to a lesser degree, showing less positive skew, and they rarely, if ever, exhibit a ceiling effect. These facts suggest that there is a broader range of skills involved in reading a passage than is normally identified and measured in multiple-choice comprehension tests, and that cloze tests measure both those that are so simple and easy that they fall below the lower limits of multiple-choice tests and those that are so complex and difficult that they fall above the upper limits of the multiple-choice tests.

But regardless of what theoretical interpretation is placed on them, these facts suggested that regressions between the two kinds of tests would be nonlinear when a broad range of scores is obtained on the cloze tests. That is, it could be anticipated that, in the lower range of cloze scores, information gain scores would show little increase as cloze scores increased. Some of the items in the cloze tests are easier than those in the multiple-choice tests; and so the multiple-choice tests would fail to discriminate among students scoring in the low range of the cloze tests, and the slope of the regression would be zero. When students' scores fall into a somewhat higher range on the cloze tests, however, increases in cloze scores should be accompanied by increases in gain scores, since the two tests are presumably both discriminating in this range. Finally, when students' scores fall in the high ranges on a cloze test, increases in cloze score should not be accompanied by further increases in information gain since the ceiling of the multiple-choice tests has been reached and no further increases in those scores are possible, even though the scores on the cloze tests may continue to increase.

It was also anticipated that one other effect might influence the shape of these regression curves. The pre-reading tests used to calculate gain scores are quite difficult and scores on them might continue to rise as cloze scores reach very high levels. If so, the gain scores could be expected to decrease as cloze scores reached these very high levels. This would occur because the gain score is obtained by subtracting the student's pre-reading score on a multiple-choice test from his post-reading score. Thus as cloze scores increase, larger and larger amounts would be subtracted from the post-reading scores, which are remaining at the constant level determined by their test ceiling, thereby yielding smaller and smaller gain scores. Coleman (1968) observed an effect that might have arisen in this way. He determined the mean cloze scores for a number of passages and also obtained a mean gain score for each passage using a modified version of the cloze procedure. When he plotted mean gain scores as a function of mean cloze scores, the gain scores rose as the mean cloze scores rose until the cloze scores reached a fairly high level, and then the gain scores declined steadily with further increases in mean cloze scores.

Procedures

Testing Design: A somewhat elaborate procedure was used to avoid the contamination that would result from giving both a cloze and a comprehension test over the same passage. First, the students were tested in order to form pairs matched for reading ability. Second, one pair member was given a cloze test made from a passage in order to determine the pair's cloze score on the passage. Third, the other pair member was given a comprehension test over the passage before he had read it in order to measure how much prior knowledge the pair had of the passage. Fourth, approximately eight days later he was asked to read the passage

and immediately retake the test. The difference between the two scores on the comprehension test was taken as a raw measure of the information gained by the pair. Subsequent analyses then examined the regression of raw and residual gain scores on the cloze scores.

Tests: The test used to match the pairs of students was a cloze test made from a 263-word passage drawn from a psychology textbook by Kretch and Crutchfield (1958). The test contained 52 items made, administered, and scored by the version of the cloze procedure used throughout these studies. In this procedure, every fifth word is deleted from the passage and replaced with an underlined blank of a standard length. The students are told to write in each blank the word they think has been deleted, and their responses are scored correct when they exactly match the word deleted, disregarding obvious spelling errors. No time limits are imposed on taking the tests. To match the students, they were first ranked in the order of their scores on this test and then, taken in order a pair at a time, one member was randomly assigned to group X and the other to group Y.

The cloze and the multiple-choice tests used in the study were made from two passages, designated passages A and B, drawn from the same source as the passage used to make the matching test. Passage A was 469 words in length and passage B contained 398 words. Each represented a fairly elementary description of a psychological experiment. Five forms were made of the cloze test over each passage, the first form by deleting words 1, 6, 11, etc., the second form by deleting words 2, 7, 12, etc., and so on. These different forms were randomly assigned to the subjects. The reason for making these five forms is that a cloze test actually represents only a sample of the cloze items that can be made for a passage. By using all five forms, this source of sampling error can be eliminated. The cloze scores were expressed as percentage scores.

A 34-item multiple-choice test was made for passage A and a 39-item test for passage B. Each item had four alternative responses. At the time this study was conducted, little work had been done on item definitions, and so only limited constraints could be put on the writing of the items included in these tests. The constraints that were imposed were (a) that every item that seemed possible be written, (b) that the wording of the questions and the responses adhere as closely as possible to the wording of the passage, (c) that, when one item's stem cued the response to another item, one be selected randomly for the test and the other be deleted, and (d) that a process taxonomy (Bloom, 1956) be used to represent as broad a range of question types as possible. The questions were tried out on several students for the purpose of identifying ambiguous and illogical items. Such items were revised, but no item was revised or discarded for any other reason. The scores on these tests were corrected for guessing and converted to percentage scores.

The reliability of each test was obtained from the data reported in this study by correlating scores based on odd- and even-numbered items in the tests and correcting the correlations using the Spearman-Brown prophecy formula. The reliability of the matching test was .83. The reliabilities of the multiple-choice tests were .84 and .86 for the tests made from passages A and B, respectively. These reliabilities were based on the student's scores made after they had read the passage. The reliabilities of the cloze tests were calculated in the same manner, but by pooling the scores of all five test forms. The reliabilities were .92 and .89 for the tests made from passages A and B, respectively.

Testing: In this kind of study it is necessary to observe the regression throughout as much of the range of the two measures as possible. In order to obtain this range using a fairly small number of students, a broad age range of students was used. A total of 130 pairs was tested, 25 pairs in grade 3, 23 pairs in grade 5, 15 pairs in grade 7, 28 pairs in grade 11, 24 pairs in their second year of college, and 15 pairs in graduate school. Because of absences, the data actually obtained represent 130 and 127 pairs for passages A and B, respectively. The tests were administered in three sessions as shown in Table 1. All students

TABLE 1

Testing Sequence

Pair Member	Multiple-Choice Pre-Reading	Cloze	Multiple-Choice Post-Reading
X	passage A	passage B	passage A
Y	passage B	passage A	passage B

were given the matching test during the first period. About a week later they were given a pre-reading multiple-choice test and a cloze test. At these two testing sessions they were told that this was a study to find out how well students could guess on tests of various kinds. The post-reading test was given slightly over a week later. No time limits were imposed on the tests. This testing arrangement made it possible to obtain for each pair of students a cloze and a gain score on each passage. For example, the pair's cloze score on passage A came from pair member X's scores on the two administrations of the multiple-choice test.

Results

Regression of Raw Gain Scores: The shapes of the regressions for the two passages can be seen in Table 2. This table shows the mean raw

TABLE 2

Mean Raw Gain Percentage Scores of Pairs Falling in Successive Intervals of Cloze Scores

Cloze Score	Passage A		Passage B	
	N	$\bar{X}$	N	$\bar{X}$
0 - 4	7	2.2	3	2.4
5 - 9	9	12.3	4	12.1
10 - 14	10	8.2	9	14.9
15 - 19	10	.5	9	10.0
20 - 24	8	10.7	5	17.8
25 - 29	3	24.3	19	21.8
30 - 34	13	30.4	11	38.9
35 - 39	5	18.1	14	44.2
40 - 44	5	47.4	9	42.9
45 - 49	15	40.7	10	45.1
50 - 54	16	42.2	20	43.9
55 - 59	12	42.6	6	51.9
60 - 64	12	44.2	4	35.4

gain scores exhibited by pairs whose cloze scores fell into successive five-point intervals on the cloze scales. All scores are expressed as percentages. Values were omitted from the table for those intervals in which there was no more than one value for one of the passages. The gain score means differed at the .05 level of confidence only in the cloze score intervals 15 through 19 and 35 through 39 percent. Moreover, neither regression fell consistently above the other throughout the range in which comparisons were possible. These two facts made it seem likely that, for at least these tests, the regressions of information gain on cloze scores exhibit similar regressions; that is, regressions that are approximately identical in their parameters.

Regression of Corrected Scores: Two adjustments were made to correct the scores for statistical biases before they were combined and analyzed to estimate the parameters of this regression. The first

correction consisted in calculating residual gain scores. Raw gain scores yield biased estimates (Harris, 1963) of true gain scores because, in part, errors in the pre-test and gain scores are necessarily correlated. For example, if a student should make a few correct guesses on the pre-test, the effect would be to reduce spuriously the size of his gain score, assuming that errors of measurement on the pre-test and post-test are independent. Bias also arises from the fact that only a part of the difference between pre-test and post-test scores represents a gain in the function measured by the post-test. A pre-test presents the student with a task that differs materially from the task represented by the post-test. The former, for example, might be described as testing inferential and other skills as well as prior knowledge of the content of the passage, while the latter might reasonably be described as depending much less on the inferential skills and other skills. If raw pre-test scores were subtracted from the post-test scores, the result would be to underestimate how much information the student gained from reading the passage. This results from subtracting from his post-test score the portion of his pre-test score that resulted from inferential and other skills rather than subtracting just that portion that represented his prior knowledge of the content of the passage. The procedures used to correct the gain scores used throughout these studies were those described by Cronbach (1970). In essence, they have the effect of calculating just that portion of a pre-test score that correlates with the post-test score and then further correcting this quantity for error of measurement before subtracting it from the post-test score.

The second adjustment was made to convert the cloze scores to their true scores. The regression models used in these studies do not take into account the error of measurement in the independent variable, the cloze scores. This error of measurement causes a student's observed cloze score to lie farther from the mean than its true value, the mean of the values that would be observed if he could be administered the test an infinite number of times. The effect of this bias would generally be to associate a gain score with an observed cloze score that lay farther from the cloze mean than the true cloze score with which it is associated, thus stretching the curves at the extremes. Since this bias would influence the level at which a performance criterion was placed, it was removed by correcting the cloze scores to their true scores.

Form of the Regression: In order to observe the shape of this regression, the data from the two passages were combined and a polynomial curve fitted to the regression. In this analysis a matrix of correlations was calculated among information gain and each of the variables that represent several successive powers of the cloze scores--their first power, their quadratic power, and so on up through the eighth power. A series of multiple regressions was then calculated entering successively higher powers of the cloze scores until a criterion of fit was reached.

Finding a suitable criterion of fit proved difficult. A criterion of statistical significance could not be used. In the initial effort to fit these data a stepwise regression procedure was used. In this procedure the power of cloze having the highest zero-order correlation with gain is selected as the first independent variable to be entered in the regression. Partial correlations are then calculated between the variables representing the remaining powers and the gain scores, and the one having the highest partial correlation is the next one entered into the regression. A variable in the equation was removed if at any time its partial correlation fell below the .05 level of significance, and the procedure was stopped when none of the variables not yet entered into the equation had a significant partial correlation with information gain. Only the first and cubic powers of cloze entered this equation. The curve resulting from this analysis was then plotted along with the column means obtained for the corrected gain scores. This curve fit the column means well only in the central region of the distribution where there were substantial numbers of cases. The fit became increasingly worse as the cloze scores departed from the mean, passing as many as four standard errors from three of the column means and passing at least two standard errors from the means of six columns. As a result, the customary test of significance was rejected as resulting in a systematically biased fit of the regression.

Instead, a visual criterion was used. The results are shown in Figure 1. Successively higher powers of the cloze scores were cumulatively entered into the equation until the curve appeared to pass through the central region of the column means at all levels of cloze and yet to be relatively free of variations due to the local anomalies that appeared in the column means. Eight powers of the cloze scores had to be entered into the regression equation before these criteria appeared to have been satisfactorily met. All eight of the curves were plotted along with the column means and standard deviations in order to provide an illustration of the behavior of the curves as each successive power of the cloze scores was introduced into the equation. The cross bars on the vertical lines represent the standard deviations of the scores in the columns. The numerals represent the number of pairs falling into that column.

The figure shows that each successively higher power of the cloze score removed some of the bias from the curve. The added accuracy in this respect, however, is sometimes achieved only at some cost in the standard error of the regression. With each successive parameter introduced into the equation, an additional degree of freedom is taken away from the denominator for calculating the error or the mean squares for variation about the regression line, while the parameters introduced subtract decreasing amounts of variance from the numerator. Consequently, while each variable added always increases the size of the multiple correlation, at some point adding new variables must begin to increase the size of the standard error of the regression. In the case of this regression, this effect occurred with the introduction of the eighth power of the cloze scores. It can be seen from the figure that adding this

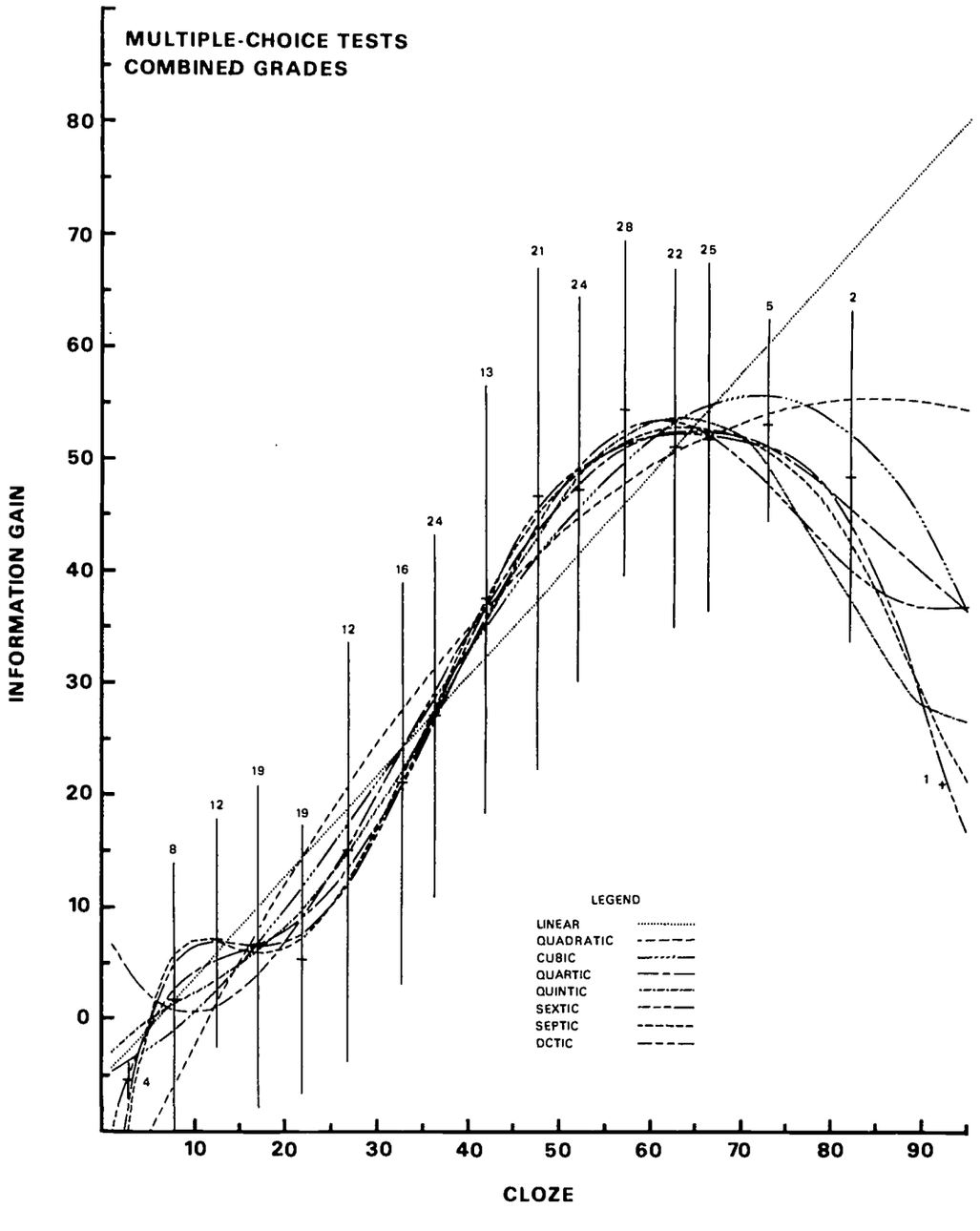


FIGURE 1. Successive polynomial fits to the regression of residual information gain scores on cloze true scores. The cross bars on the vertical lines represent the means of the gain scores for cloze scores in each five-point interval. The vertical lines represent the standard deviations of those scores, and the numerals below each bar show the number of pairs in that interval.

term resulted in a better fit to the column means. However, the mean squares due to error rose from 2.622, which was associated with the seventh degree polynomial, to 2.631 and continued to rise thereafter.

The shape of this regression curve conformed fairly well with expectations based on previous experiences with the two kinds of tests. The general shape of the curve was roughly similar to a normal ogive curve--low and flat at the lower range of scores, rising sharply in the intermediate range, and flattening out at a high level in the upper range of cloze scores. There also seemed to be some slight tendency for the curve to turn downward in the region of very high cloze scores, but not for the expected reason. A plot of the pre-test scores showed that they too tended to show little additional increase above the intermediate range of cloze scores. In addition, when the scores were inspected for the pairs whose cloze scores fell above 70 percent, none had a post-test score of more than 80 percent. However, the cases in this region were too sparse to bear much interpretation.

An unexpected but potentially important effect appeared when scatter plots were made of the scores obtained by students at various grade levels. While the curve for each age of student appeared to roughly parallel the others in shape, there also appeared to be some separation between the curves, with the older students obtaining higher gain scores at every cloze score. The reality of this impression could not be tested statistically, however, since the overlap in the ranges of scores was slight at the lower age levels where the effect appeared to be most prominent.

Discussion

The chief result of this study was to suggest that it might be feasible to identify a passage performance criterion using information gain as a part of its rationale. The fact that the two curves were nearly identical in their parameters indicated that the methods used to construct the comprehension tests were likely to produce tests having at least a useful degree of conformity to the instruction and to measure a fairly uniform set of behaviors across passages. The confidence in these judgments was further increased by the fact that the general shape of the regression could be successfully predicted from previous experience with other tests of the same types. Consequently, it seemed worthwhile to pursue the studies further.

Two unanticipated results seemed particularly useful to investigate further. The tendency for older students to show more information gain at each cloze level suggested that the regressions might differ in several ways at various grade levels. If this were true, it might make it necessary to derive a different performance criterion for students at each grade level. The second result was the fact that the downward trend of gain score in the upper range of cloze scores, if one existed

at all, could not be explained as merely the mathematical consequence of the pre-test and post-test regressions having shapes that were not exactly parallel. This prompted speculation that perhaps students for whom passages are very easy might be bored by the passages. Thus, any downward trend that might occur might ultimately demand an explanation that took attentional variables such as interest into account. And the pursuit of this speculation eventually led to the formal incorporation of interest variables in the criterion selection model.

STUDY II

Purpose

Before a major study was undertaken it seemed desirable to do more exploratory work, since the identity of the regressions of information gain on cloze scores had been demonstrated on only two passages. Also, the method used in the first study to measure passage comprehension did not permit a satisfactory degree of control over the item writing and, therefore, over the instructional conformity and behavioral consistency of the tests; the effects of a student's grade level on the regressions needed to be examined; it appeared that a student's interest in the reading task might also need to be taken into account when a performance criterion is identified; and, finally, the design of the previous study was cumbersome and expensive relative to the amount of data it yielded. The second study was designed to explore each of these problems and to provide independent verification for the study to follow it.

Generalizability of the Regression Identities: The initial study was concerned with the question of whether a regression identity might occur for any passages, and so that study used two passages that were similar in difficulty in order to assure some chance of observing an identity if one could occur. However, identifying a passage performance criterion implies that this criterion is equally applicable to any of the passages within some specified population of passages. Thus, it was necessary next to determine if regression identities occur for large numbers of passages, and to explore the possibility that the parameters of the regressions might differ systematically along some dimension of passages.

Passage difficulty appeared to be a dimension along which regression curves might differ. A reasonably credible argument could be offered to justify the expectation that regression identities might be observed if the passages are highly similar in difficulty. The shape of a regression between two tests is generated by the relationship of the respective test metrics to each other, that is by the relationships existing between the joint distributions of the item difficulties and item discrimination indices in each of the two tests. So presumably when regression identities occur it is because each of the test-making procedures produces tests having uniform metrics across passages. In tests made from passages of highly similar difficulty this might occur because such passages contain highly similar language. Readability formulas use variables based on the linguistic features of passages to predict passage difficulty and the modern formulas (Bormuth, 1969a) predict approximately 84 percent of the total observed variance of passage difficulty and 95 percent of the reliable variance. Moreover, the linguistic variables on which these predictions are based are fairly highly correlated. Thus, passages that are similar in difficulty tend to be homogeneous in the type of language structures they contain.

The items made by both the cloze and the question procedures closely conform to the language structures in the passages from which they are made. If two passages contain many identical structures such as adjectives, or conditional clauses, cloze items made from the two passages are likely to bear similar relationships to structures of the same types. And such items tend to be homogeneous in at least the metrical property of difficulty. Coleman (1968) found, for example, that items deleting the same parts of speech tend to exhibit a considerable amount of homogeneity of difficulty across passages. Similarly, items made by the question-writing procedure conform to the language in the passages since they are formed by transforming language structures in the passages (Bormuth, 1970b). For example the question How are items made by the question-writing procedure formed? can be shown to have been derived by performing a set of transformations on the syntactic structure of the preceding sentence. Moreover, items derived from the same type of syntactic structures exhibit homogeneity of item difficulties across passages (Bormuth, Manning, Carr, and Pearson, 1970). Thus, given that the passages are of relatively similar difficulties and that a uniform procedure is used to draw the items from different passages, the distributions of item difficulties should be similar for all passages.

In summary, then, it seemed reasonable to expect regression identities to exist among passages of similar difficulties because (a) such passages tend to contain a similar type of language, (b) the items in question and cloze types of tests conform to the linguistic structures in the passages from which they are made, (c) items made from similar linguistic structures exhibit similar item difficulties, and thus (d) each type of test-making procedure should produce items having a distribution of item difficulties that is fairly uniform across passages. And the relationship for these two types of tests of at least this determinant of the metrics should be approximately the same for all passages having the same level of difficulty.

But it should also be noted that this argument provides no strong reason for expecting regression identities to hold across passages at different levels of passage difficulty. The frequencies of various types of language structures vary systematically with passage difficulty. This might or might not alter the metrics of each type of test, depending on the effects of passage difficulty on the joint distributions of item discrimination indices and item difficulties of the tests. It is tautological that there would be an effect on the difficulties of the items in the tests, since passage difficulty is normally defined as the mean of the item difficulties. However, there is little or no evidence that would suggest whether the relationship between test metrics would or would not change in other ways as a function of passage difficulty. At most it could be argued from present knowledge of language structure that any change in the regression would probably be systematic along the difficulty dimension, provided that any change occurred at all.

Control of Item Writing: It is essential that the tests used to observe regression identities have an operational and replicable conformity to the passages from which they are made. Identifying a passage performance criterion implies that a regression of a particular form holds for all pairs of cloze and question types of tests made from the same passages. And it implies that this regression is representative of all the responses of these types to the passage and not merely an artifactual and trivial effect of the way a particular test maker chose to make the tests. Artifactual and trivial regression identities can be produced almost at will by conventional test-making procedures. The test maker writes a number of items, selecting and altering their phrasing until he has produced whatever item difficulty and discrimination indices he wants, and then he selects for inclusion in the tests those items that produce the test metrics necessary to obtain whatever regression he has chosen to obtain. These regressions, however, are artifactual in the sense that they are not representative of all the responses of those types evoked by the passage, but are merely the results of the test maker's machinations. And they are trivial in the sense that they cannot be interpreted as showing the values of the response levels on the cloze tests and as properties of the passages, but only as one of many arbitrary regressions that could be produced in the same way.

In order to regard regression identities as nontrivial, it is necessary for the tests to be derived in a manner that permits the tests and their regressions to be regarded as a property of the passages. To be specific, the test items made from a passage should be operationally and replicably derived as specified functions of the content and language of each passage, so that the items that are written are wholly determined by the test-writing rules and the passage--and not by the judgment of the test writer. Also, an unbiased sampling operation should be defined to draw those items actually included in the tests if, as is generally the case, it is impossible to include all the items that can be derived. If this is done, the test items may be regarded as a property of the passage with no other factor systematically influencing the properties of the tests, and so the properties of these tests, including their regressions, may also be regarded as properties of the passages. This is not to say, however, that a test must contain every conceivable type of item that could possibly be made from a passage. Quite the contrary, some item definitions generate items that are not particularly useful. Rather, it is to say that once a particular type of item has been identified as being one of those to be included in the tests, no item of that type should be excluded except by the sampling operations of the test. If items are excluded for any other reason, then the test is no longer a property of the passage but is also a property of the test writer's idiosyncrasies.

In the preceding study an effort was made to achieve this goal by laying down some specifications for the writing and selection of the test items. However, these measures fell short of the amount of control that is desirable. The alternative response choices represented a

particularly difficult problem. On the one hand, it was desirable to avoid introducing extraneous influences on the difficulty of the item by including response alternatives containing vocabulary and syntactic structures that were more difficult than the language in the passage itself while, on the other hand, it was difficult to find incorrect alternatives in the passage that might reasonably attract responses. In order to avoid questions that could be answered by obvious eliminations, the constraints on the item writing were often relaxed. A different problem arose because alternative responses may or may not be conceptually related to each other in various ways. For example, the conceptually related set animal, mammal, dog, and collie might constitute the alternatives for a question while dog, bird, snake, and fly might also serve for the same question. Choosing either to use or not to use related alternatives and choosing the particular form of relationship for a set probably affects the nature of the task, but no way had been developed to control and justify these decisions. Since most question stems are somewhat mechanical transformations on the passage, only minor problems occurred in the wording of the question stems; among the problems that did occur were those questions with apparently alternative phrases such as What kind of, What color of, or What hue of. Rather, the chief problem in writing question stems occurred with efforts to enumerate every question that could be written so that the tests could be regarded as a sample drawn from that population.

As a consequence of these problems, the results of the preceding study were subject to reasonable alternative interpretations, the most damaging one being that the shapes of those two regressions were determined by unrecognized biases in the test construction procedure and therefore that the apparent identity of regressions was merely fortuitous. Since the test construction rules were vague, there is no decisive way to refute this interpretation.

Several measures were taken in this second study to more closely approximate the goal of constructing comprehension tests that have an operational and replicable conformity to the passages. The use of multiple-choice tests was abandoned in favor of the use of short-answer completion tests. The reason for using response alternatives had been to avoid the unreliability that seemed to be involved in evaluating the correctness of the various responses that students write for completion questions. However, because of the difficulties of identifying alternative responses and composing sets of responses, any gains in scoring reliability seemed to be more than offset by losses in the passage conformity of the tests. Promising procedures are being developed to handle this problem (Schlesinger, 1970, and Guttman and Schlesinger, 1967), but they must be more rigorously defined and extended in range before they will represent an improvement over the completion question. In the meantime an objective procedure for scoring completion responses had been developed for use in another study (Bormuth et al, 1970), and so it was decided to use short-answer completion questions in this study. Also, somewhat more adequate methods were devised for enumerating some of the populations of items that can be derived from a passage.

Grade Level Effects: In order to use the same performance criterion with different groups of students it is necessary to assume that their scores would exhibit identical regressions if they were included in an experiment of the type reported here. If the regressions for different groups differed in shape and if those regressions were subsequently entered into the criterion selection model, the model would ordinarily yield a different performance criterion for each group. Thus, it was a matter of considerable interest when the scatter plots in the previous study suggested that the regressions might vary with the grade level of the students. These effects appeared even more clearly in one of the field trials of some of the materials used in the present study. A fourth- and a seventh-grade group yielded regression curves that appeared to diverge somewhat, and the seventh-grade students showed higher gain scores at every level in the range through which their cloze scores overlapped. The explanation of this phenomenon was obscure at that time and remains highly tentative. However, it seemed fairly likely that a single performance criterion could not apply equally to students at all grade levels and that this study should be designed to verify that possibility.

Interest: By the time this point in these studies had been reached, it had become evident that a model for identifying a performance criterion should also include a measure of the students' attitude toward the materials, and the rationale described in the presentation of the model was constructed to state this justification. The theory relating stimulus complexity to attentional behaviors and preference choices has been developed sufficiently to make it of some use in developing policy-making models of this sort. (See Dember and Earl, 1957; Berlyne, 1960; and Dember, 1965, for discussions of the theory in this area.)

This theory, often referred to as the Dember-Earl theory, regards different tasks or objects as being comparable in terms of a scale of complexity, where complexity refers to the amount of information, in the information theory sense, contained in the task. The theory claims first that gaining information has a positive motivating effect that controls perseverance of attention and preference choices. Next, it claims that each individual has, at a given moment in time, a complexity value relative to a given class of stimuli, this complexity value being the individual's momentarily preferred level of complexity. Thus, given a series of objects or tasks of the same class arrayed on a scale of complexity and a person's preference ratings of those tasks, the plot of their preference ratings of the tasks as a function of their complexities would exhibit an inverted V-shaped curve and the apex of this curve would define that person's complexity value. The theory then asserts that shifts in a person's complexity level are toward tasks and objects at a slightly higher level, toward tasks referred to as pacers. The person exhibibits the greatest amount of attentional perseverance to pacer tasks. What makes this theory useful for making policy decisions is not that the information content of the stimulus is sufficient to account for all of the variance in indices of interest. Indeed, it seems probable that there are other major sources of variation not yet taken into account. But rather, its utility lies in the fact that

it conceptualizes interest as being systematically influenced by dimensions of the stimulus rather than as being merely some internal state that is idiosyncratically related to the student's experience and to the unique combinations of features of the stimulus.

Kamman (1966), for example, operationalized these concepts for the purpose of studying preference for poetry as a function of the complexity of poems. He determined the students' cloze scores on a series of poems; had them rate each poem with respect to how well they would like (a) to hear a professor discuss it, (b) to memorize it, (c) to discuss it with a friend, and (d) to debate it; and then plotted these preference ratings against the relative cloze difficulties of the poems. Each of the plotted curves followed roughly the inverted V-shape curve predicted by the Dember-Earl theory. A result of major methodological interest for the present studies was the fact that the curves peaked at considerably different levels of cloze performance, depending on the use for which the students were rating the poems. This suggested to Kamman that dimensions of the social context in which the poems were to be used had important effects on the regressions, and his interpretation was supported by the fact that a measure of the students' tendency to prefer the easier poems correlated positively with a measure of debilitating anxiety and negatively with a measure of facilitating anxiety. Thus, it seems essential in establishing performance criteria to make the measures of preference as specific as possible about the use for which the materials are being rated.

Procedure

It is necessary to use large numbers of students in order to obtain stable estimates of the regression curves over a wide range of scores on each type of test. In order to accomplish this with reasonable economy, it was necessary to design a more efficient testing procedure than the one used in the preceding study. Using matched pairs of students was costly in terms of the amount of clerical time required to account for materials, the amount of testing time wasted in passing out materials, and the large losses of cases due to absences at one or more of the testing sessions. In the present study a design was used that permitted all data to be gathered in a single testing session. This was achieved by giving the student a cloze test over one passage and a preference rating scale and pre- and post-comprehension tests made from a second passage of matched difficulty.

Passages: A total of 44 short passages taken from the four forms of the Gray Oral Reading Paragraphs (1963 edition) were used in this study. Each form of this test contains 13 passages that are scaled in levels of difficulty so that they range from pre-primer levels to paragraphs sufficiently difficult to challenge the most able reader. The two easiest paragraphs in each form were dropped because they were too short and devoid of content to provide useful measures of comprehension.

These paragraphs were used because the test makers had matched the paragraphs for word recognition difficulty at each level. Also, the author (Bormuth, 1968) had used them in an earlier study in which their cloze difficulties were obtained, and this permitted some of the paragraphs to be switched from one difficulty level to another in order to obtain still closer matches of the passages within each difficulty level. The passages dealt with a variety of topics. There was, however, a definite tendency for topic and writing style to correlate with difficulty. The easier passages dealt with topics typical of the reading books traditionally used with lower grade children and the style of the writing, to some extent, "talked down" to the reader.

These passages were less than ideal for this reason, and also because they were so short that it was difficult to obtain tests long enough to provide high reliability. Their chief virtues were that they had known difficulty levels, which permitted them to be matched, and part of the tests over them had been made for an earlier study. Preliminary trials indicated that they would have sufficient reliability for the purpose of this study.

Completion Test Construction: A ten-item test was made for each passage using wh- rote questions that require the student to write short answers. All of the questions were derived either from the syntactic structures within sentences or from the anaphoric structures that sometimes connect sentences. Since a fairly detailed description of this question-writing procedure is available elsewhere (Bormuth, 1970b), it will be described only briefly here. Consider sentence (12), for example.

(12) Joe gathered his tools and packed them neatly.

(13) Then he went home.

(14) Who gathered his tools and packed them neatly?

(15) Who gathered his tools?

(16) Who gathered tools?

(17) Whose tools were gathered?

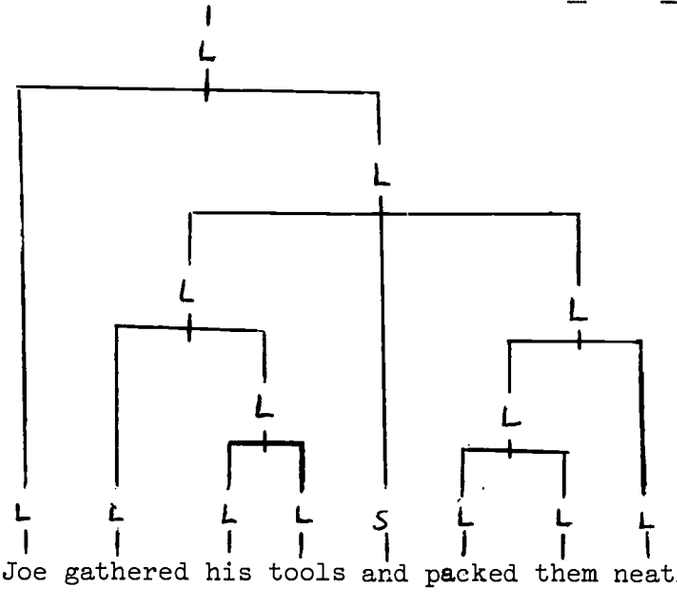
(18) When did Joe go home?

(19) Who went home?

It is possible go derive a fairly large number of wh- questions from this sentence by first analyzing its syntactic structure and then performing manipulations on that structure. Question (14) was derived by (a) deleting Joe, (b) replacing it with the wh- pro word, who, and (c) inserting a question mark at the end of the sentence. Wh- pro words consist of words such as who, what, where, and when and occasionally of phrases such as

what kind of, how many, at what time, and for what reason. Question (15) was formed in the same way but with the added step of deleting one of the two predicates of sentence (12). Questions (16) and (17) were derived after the still further deletion of his. Questions (18) and (19), on the other hand, were derived from the anaphoric structures that connect sentences (12) and (13). An anaphoric structure consists of an anaphora and an antecedent. The anaphora is a pro word or phrase that stands for an antecedent that occurs earlier (or, in a few cases, later) in the passage. The word he in sentence (13) proes the noun Joe, the word then proes the whole of sentence (12), and the word his in sentence (12) proes Joe within the same sentence. An anaphoric question is obtained any time a wh- pro word replaces an anaphora. Thus, questions (17) and (18) would also be considered anaphoric questions.

The transformation rules by which these questions are derived perform manipulations on the phrase structures of sentences. An example of such a phrase structure is diagrammed in (20). The L and S's in this diagram



(20) Joe gathered his tools and packed them neatly.

are called nodes. Normally nodes are labeled with symbols such as NP, V, PREP, and so on to show that the constituent under it is a noun phrase, verb, or preposition. In this case each is labeled only with L and S to show whether it is a lexical or structural element. Lexical constituents are, roughly speaking, nouns, verbs, adjectives, and adverbs or phrases containing one or more of those parts of speech. The symbol standing for a word or phrase is called a node and the lines below it leading to the right and left are each called branches from the node. The question transformations are written as formulas that state how the nodes in the phrase structure, when conventionally labeled, are manipulated to obtain a question.

The items written for the tests were selected using a randomized, sequential procedure. First, the grammatical structure of each sentence was diagrammed, the lexical nodes identified, and a number assigned to each lexical node. Second, one of the numbered nodes was drawn randomly and a question formed by replacing the branch that represented the shortest constituent with an appropriate wh- word and then performing the remainder of the transformations necessary to form a grammatical question. This operation might, for example, produce a question such as (14). Third, the question was then inspected to identify constituents that could be deleted without making the question incomprehensible. For example, his could be deleted from (15), but tools could not be deleted from (16). Fourth, one of these deletable constituents was then randomly selected and replaced with an appropriate wh- pro to form an additional question from the same question stem. This step would produce questions such as (15). Finally, this process was continued until terminal questions were reached. These were questions such as (16) and (17) in which no further deletions were possible without introducing more than one pro into a single question. The items included in the tests were randomly drawn from these terminal questions.

The immediate purposes for adopting these rather elaborate item-writing rules were to insure that the tests had good instructional conformity and behavioral consistency. The rules were designed to exclude, as much as possible, any effects that might arise out of the idiosyncrasies of the test writers. The sequential procedure was developed for two reasons. First, when the entire sentence is included in the question, the question is often grammatically awkward and confusing to read. Second, the passages were so short that it would have been impossible to obtain tests long enough to be reasonably reliable had only one question been made per sentence. And when more than one question is made per sentence, a sequential procedure of some sort must be used since the selection of one node for questioning determines what other nodes can be questioned. Finally, the ultimate purpose of closely controlling the test-making procedure was to meet the most fundamental criterion required of all operations that are intended to produce results that are acceptable for use in supporting either scientific or public policy statements. That is, that the results must be operationally replicable. Obviously, results cannot be depended on if they are based on tests in which the test writers have exercised a large measure of judgment in ways that are essentially unknown.

Although these tests achieved a fair degree of operational replicability, they were not entirely satisfactory for several reasons. First, these procedures do not always produce grammatically acceptable questions. In part, these anomalous questions were due to the fact that the processes by which constituents may be deleted within questions are not yet a fully understood aspect of grammar. And, in part, it is also due to the fact that transformational grammar was, at the time these tests were written, based on vaguely defined concepts of deep structure. In either event,

the test writer is forced either to accept a few questions that seem grammatically questionable or to rely on his judgment, and for those questions it seems doubtful that different test writers could perfectly replicate each other's work. Research in progress by Patrick Finn, one of the author's graduate assistants on this project, apparently shows that nearly all of the problems of this type may be solvable by basing the transformation on a case structure analysis of the sentences. But since this work is still in progress, this criticism holds for all of the studies so far conducted in this series.

The wh- pros inserted to form the questions presented the other major problem. It was not always possible to use a one-word pro to form a question. And when more than a one-word pro is required, the pro phrase that is used contains a pro word along with a word or phrase that names an abstract category that includes the concept referred to by the constituent replaced by the pro phrase. For example, if one attempted to write a question in which the pro replaced the word green in sentence (21), he might obtain (22), which is semantically incorrect.

(21) The boy ate the green apples from the tree.

(22) *Which apples from the tree did the boy eat?[1]

(23) What (color/hue) were the apples the boy ate from the tree?

Its unacceptability may be seen by attempting to question the underlying sentence, The apples were green, in the same way by asking *Which were the apples?. Consequently, at this point the test writer is forced to use a pro phrase that contains a pro word, usually what, plus a word referring to a concept that is more inclusive than the concept the pro replaces. In (23) the words color or hue serve that function. The problem here arises that no rules have yet been suggested for enumerating each of the alternative words that could be used in a given question and then for regularly selecting just one. The test writers on this project employed their judgments in these matters, following only the rather vague rule that the abstract word should be both as general as possible yet as common as possible, where commonness was determined by its frequency in the Thorndike and Lorge (1944) counts. Consequently, on these grounds it also can be seen that the tests, and therefore the results, in these studies may not be completely replicable. However, the effects of these remarks should be realistically moderated by pointing out that the nonreplicable items constituted a fairly small proportion of the items in any test and even the nonreplicable items can be identified replicably.

[1] An asterisk before an expression indicates that it is an unacceptable form.

Completion Test Scoring: The use of completion items has often been avoided because the responses seemed difficult to score reliably. The correctness of responses frequently appeared to be purely a matter of the subjective opinions of test scorers rather than a type of issue that could be decided by rules. This view has turned out to be largely incorrect: Scoring rules that were first developed (largely with the help of David Pearson of the University of Minnesota) for use in another study (Bormuth et al, 1970) have been further developed in the present series of studies, and these rules reduce the role of subjective judgments to negligible proportions. Since this scoring system will be the subject of another report, only its main features will be described here.

The generic response to an item is the phrase replaced by the wh-pro to form the question. Thus, given sentence (12) and question (14), the generic response is Joe since it was the constituent replaced by the wh- pro. All observed responses are scored correct or incorrect depending on the nature of their relationships to the generic response.

(12) Joe gathered his tools and packed them neatly.

(13) Then he went home.

(14) Who gathered his tools and packed them neatly?

(24) What did Joe do?

(25) gathered his tools and packed them neatly

(26) gathered and packed his tools

(27) gathered

An observed response may be either formally or semantically related to the generic response. If the observed response is formally identical to the generic response, it is scored correct. For example, Joe would be regarded as formally identical with respect to question (14), and response (25) would be regarded as formally identical with respect to (24). However, some observed responses are formally identical to only a part of the generic response; responses (26) and (27), for example. The general rule followed in these cases was that only the word that was grammatically the head of the generic response was required in order to score the response correct. However, this general rule does not apply in some types of grammatical structures. When the generic response was a coordinated structure, only one of the coordinated structures was required in order to score the response correct. Thus, (27) would be regarded as correct. Transformations provided a major device for determining formal relationships. Thus, since (25) can be transformed to obtain (26) without losing its head words, (26) would be regarded as formally related to (25) even though its superficial form appears to be different.

Some correct responses, such as (28) and (29), bear no formal relationship to the generic response. In some cases, as in (28), these are constituents from another region of the passage, and they have one of

(28) went home

(29) brought his instruments together

(30) *brought his wrenches together

the case relationships to the question stem that is allowed by the wh-pro. The pro in (24), for example, is what...do, where do replaces the verb phrases that are being tested in (12). However, do could also replace the verb phrases in (13) and it would modify the same subject, Joe. Responses of this type were regarded as correct. On the other hand, some observed responses, such as (29), bear synonymity relationships of various kinds to the generic response. For example, brought together in (29) bears what is called (Bormuth, 1970b) a symmetrical synonymity relationship with gather in (25). That is, in this sense of the meaning of gather the two forms seem to be mutually substitutable. Instruments and tools, on the other hand, bear a hierarchic synonymity where instruments refers to a general class of objects that includes the class of objects referred to by tools and is said to dominate tools in that hierarchy. If a correct formally related response was recoverable from an observed response through either of these two types of synonymity relationships, the response was scored correct. However, it should be noted that had the observed response contained a term, such as wrenches in (30), that was hierarchically dominated by the term in the generic response, this observed response would have to be scored wrong, unless, of course, the term tools was an anaphoric form with wrenches standing as its antecedent somewhere in the passage.

While this description provides only a brief sketch of just the main features of this scoring procedure, it should be sufficient to demonstrate that the correctness of responses to completion questions can, in large measure, be decided by rules that are sufficiently explicit to permit highly replicable results. Throughout these studies continual comparisons were made between the work of different test scorers independently scoring the same responses. All but a trivial number of disagreements among scorers resulted either from clerical counting errors or from the failure of one or both of the scorers to correctly apply the scoring rules.

Preference Scale: A preference scale was devised to obtain a measure of the students' willingness to study the textbook that a passage supposedly represented. This was the seven-point scale shown in Figure 2 with the printed instructions that accompanied the scale. The instructions for the scale directed the student to rate the passage with respect to a particular use, as a school textbook. For convenience in data handling, the scale scores were transformed such that -3 corresponded to 1, and 3 corresponded to 7.

This story was taken from a school textbook. How well would you like to study this textbook in school? Mark your answer by circling one of the numbers below.

-3	-2	-1	0	1	2	3
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

FIGURE 2. The scale used to measure the student's preference for the textbook that a passage supposedly represented.

The passage to be rated was printed at the top of a sheet of paper and the instructions and the scale were printed below it. Before the student read the passage these instructions were read to him along with the verbal labels below the scale, and he was told how to indicate his preference for the passage by marking a number on the scale.

Cloze Tests: Five forms of a cloze test were made from each passage. These tests were made, scored, and administered in the usual way.

Testing Design: Each test booklet contained tests over two passages that were both within the same difficulty level. The first page contained blanks for the student's name, grade, and other data useful in data handling. The second page contained the completion test over one of the two passages. The student was told that this was a study to find out how well he could guess on various kinds of tests made from short passages and that he should study the items carefully and try to guess their answers. The third page contained the cloze test made from the second passage used in the booklet. He was told how this test had been made and that he should try to guess the word taken out to form each blank and to write it in the blank. The fourth page contained the passage from which the first completion test had been made. The student was told that he was going to read this passage and then answer some questions about it but that, before he took that test, he would be asked to show how well he thought he would like to study the textbook that the passage was taken from. He was then told how to use the rating scale. When he finished rating the passage he was told that he could study the passage again before turning to the test but that he could not look back at the passage after he had turned to the test. The fifth page contained the same completion test that the student had taken before as a guessing

test. The student was urged to attempt every item on the completion and cloze tests, and he was told that although misspelling would not be counted against him, the test administrator would give him individual help by spelling any word asked of him. No time limits were imposed during the testing. The tests were all administered in a single period that ranged from 40 minutes at grade 3 to roughly 25 minutes at grade 12. This included a short break after the cloze test had been completed. The groups varied in size depending upon what was convenient in particular schools.

The test booklets were assembled using a counterbalanced design. The four passages at a difficulty level were first coded A through D. Then the six possible pairs of these passages, AB, AC, AD, BC, BC, and CD, and the reverses of those pairs, BA, CA, etc., were identified. These combinations represented each block of 12 booklets, where the first letter in each pair stood for the passage to be used for obtaining the cloze measure and the second letter stood for the passage to be used to measure information gain. Ten such blocks, or 120 booklets, were made from the passages at each difficulty level given, a total of 1320 booklets. The five forms of the cloze test over each passage were randomly assigned to the appropriate booklets in its difficulty level. Each of the ten blocks in each difficulty level was assigned to one of the stacks representing the ten grade levels to be tested. Thus there were 132 booklets in each grade level's stack when the 11 difficulty levels had been combined. The booklets in each stack were then placed in a random sequence. When the tests were administered, the test administrators took the number of tests required for the classroom in which they were testing from the top of the appropriate grade level stack and passed them out to the students starting at the top of the classroom's stack.

Students Tested: The students tested were enrolled in grades 3 through 12 in the schools of a suburban residential community where the residents' occupations were largely business and professional. A systematic bias occurred in this sample of subjects that affected the interpretation of the results. Although all students came from the same school district, the high school students were drawn from a wider and generally less able population than the other students due to the fact that the high school drew its students from a wider community. This fact was discovered only after the effects had been found in the data. A total of 132 students was tested from each grade level, with 12 students in each grade level taking the tests over the materials at each of the 11 levels of difficulty. Thus, the sample tested totaled 1320 students. It should be noted that this provided a design that was almost completely counterbalanced. That is, it provided a factorial design having 10 grade levels by 11 difficulty levels with four passages nested within each difficulty level. Any differences between passages within a difficulty level were counterbalanced by the fact that every passage was paired with equal frequency with every other passage on each of the tests. Thus, while such differences might continue to influence the scores, their

effects would average out to zero. The only departure from this counterbalancing was that the five forms of the cloze tests were counterbalanced within each difficulty level but not within each grade level, but these effects were randomized.

Test Reliabilities: The reliability of each test was calculated from split-half scores of all students who took the test. The mean of the reliabilities was .62 for the cloze tests, .44 for the information gain tests, .27 for the pre-reading completion tests, and .49 for the post-reading completion tests. The reliability of the willingness-to-read preference scale could not be estimated from these data.

Score Adjustments: In order to make unbiased estimates of the parameters of the regressions observed in these data, it would be necessary to make a number of adjustments of the scores. However, this was not done for several reasons. First, the chief purpose of this study was to determine whether regression identities could be observed under the conditions provided by the study, and this could be done without making score adjustments. Second, these adjustments are expensive to set up and compute and this study was not being directly funded. Third, the lack of funds made it necessary to employ design features that make the data less than desirable for direct use in identifying a passage performance criterion. For example, field trials had shown that giving a pre-reading and post-reading test over the same passage within the short time spans involved in this design resulted in substantially higher scores on the post-reading tests. Thus, the gain score does not represent exactly the behavior that the author thinks should be represented in a passage criterion score. Moreover, the passages on which the tests were based can hardly be claimed to be representative of instructional materials, the population of materials to which the performance criterion should be relevant. All of the cloze and completion test scores, therefore, represent percentage scores that have not been corrected.

Regression Analyses: An elaborated form of the step-wise and polynomial regression models was used in these analyses. The basic dependent variables were cloze scores (C), student grade levels (G), and passage difficulty levels (D). The polynomial terms consisted of the first three powers of these basic variables (C to C^3, G to G^3, and D to D^3). The interaction terms consisted of all possible two-way cross-products (CG to C^3G^3, CD to C^3D^3, and GD to G^3D^3) and of the three-way cross-products through the squares of the basic variables (CGD to $C^2G^2D^2$). This entire set of variables and various subsets within it were used in analyses where hypothesis testing was the main object. In the analyses where the main object was to obtain curves for plotting purposes, however, the difficulty was not used and higher powers of the cloze scores were generated. This set of variables consisted of the first eight powers of the cloze scores (C to C^8), the first three powers of grade levels (G to G^3), and all

possible cross-products (CG to C^8G^3). In all of these analyses a variable was selected for inclusion in the equation on a given step if it had the highest correlation with the dependent variable when the variables already in the equation were partialed out, and if the statistical significance of that correlation exceeded a criterion of statistical significance. Variables already selected were discarded from the equations if their partial correlations fell below the criterion of statistical significance on any step. When hypotheses were tested using this procedure, the criterion of statistical significance was set at the .05 level. However, when the procedure was used to obtain curves for plotting, the problem was run to a very low criterion and the equation was chosen on the basis of its visually judged fit to the column means and the standard errors of those means.

Results

The data were examined with respect to three general questions. The basic question, of course, was whether the regression identity assumption could be said to hold. The second question was whether student grade level and passage difficulty level systematically influenced the regressions. And the third question was what were the general shapes of the curves. It does not at the present time seem feasible to test the regression identity assumption directly. A direct test might, for example, be performed by calculating the regressions within each of the 110 cells of the design or within grade levels. A significance test might then be applied to the estimated parameters to determine if the regressions differed. The problem arises from the fact that the parameters assuredly would differ, but not necessarily because of a failure to meet the regression identity assumption. The first study and analyses of the data in the present study showed that the regressions were all curvilinear, and so the slope of one of these regression curves and the intercept extrapolated from the slope change continuously throughout the range of the cloze scores. However, since most cells contained only a limited segment of the range of cloze scores, the slopes and intercepts would reflect just that portion of the curve and would undoubtedly exhibit significant differences even if the regression identity assumption held perfectly. In other words, different cells would include different segments of the range of cloze scores. Consequently, the validity of the regression identity assumption may be supported by the fact that the regressions calculated here accounted for fairly large proportions of the variances; but the assumption, itself, was not directly tested.

Willingness-To-Study Rating: The first set of analyses attempted to determine whether student grade level and passage difficulty level affected the parameters of the regression of the willingness-to-study ratings on the cloze scores. This was done using a staged series of regression analyses. The object of the first stage was to determine the regression that could be said to be due only to the cloze scores,

pooling the data across all grade and difficulty levels. This was done by using just the first three powers of the cloze scores in the step-wise regression procedure. Line 1 of Table 3 shows the results of this analysis. The C, C^2, and C^3 terms of the cloze variable entered the equation and the multiple R was .156. This equation can be interpreted as the weighted combination of these terms that produces the regression line that best describes the regression of the ratings on the cloze scores.

The second stage was based on the reasoning that if either the grade of the student or the difficulty of the passage did not affect the parameters of this regression, the terms based on them would not enter a regression equation once the terms of the cloze variable had been entered. It is obvious, of course, that grade and difficulty would correlate with cloze scores: Students are ordinarily observed to improve in comprehension performance as they proceed through school and passage difficulty is itself defined as the mean of the cloze scores over a passage. Since both variables would, therefore, be expected to correlate with the cloze scores, they would also be expected to correlate with the ratings. However, what this second stage of the analysis attempted to determine was whether grade and difficulty had any effect on the regression of the ratings on cloze that was not simply attributable to their correlations with the cloze scores.

This stage of the analysis involved two regressions. In the first, the step-wise procedure began with a matrix of intercorrelations that included the ratings, the three cloze terms obtained from the first-stage analysis, the powers of difficulty, and the terms containing all possible two-way cross-products of the cloze and difficulty terms. The three cloze terms and the intercept were forced into the equation first to regain the same equation obtained in the first phase. And then the correlations between the ratings and the terms involving difficulty were calculated, partialing out the variance accounted for by the regression equation just calculated. When these partial correlations were examined, they ranged from roughly .0 to .13 with corresponding F values ranging from roughly .02 to 23.41, some of which could be considered statistically significant since an F of 3.85 between 1 and 1000 degrees of freedom has a probability of less than .05. The regression analysis was then completed, producing the results shown in Line 2 of the table. Only D and D^3 entered the equation, increasing the R from .155 to .221 and roughly doubling the amount of variance in the ratings accounted for by the regressions. Thus, difficulty had a statistically significant effect on the regression. However, since none of the interaction terms (i.e. the cross-products) entered the equation, this influence did not affect the shape of the curve at each difficulty level. Rather, the indication is that the shape of the regression is the same at every difficulty level with difficulty level merely adding a constant to the level of the ratings. Hence, it could be said that aside from this small constant effect on the regression curves, difficulty was highly redundant with cloze scores in this analysis.

TABLE 3

Staged Analyses of the Effects of Grade and Difficulty Level on the Regressions on Cloze Scores

	Variables Forced	Terms Finally Entering	R
		Willingness-To-Study Ratings	
1.	--	C, C^2, C^3	.156
2.	C	C^2, C^3, D, D^3	.221
3.	C	C^2, C^3, G^2, G^3, GC^2	.395
4.	C and D	G, D, D^3, G^2D	.407
5.	C and G	$C, C^2, C^3, G^2, G^3, D, D^3, GC^2$	.421
		Information Gain Scores	
6.	--	C, C^2	.481
7.	C	C, C^2, D^2C	.509
8.	C	C^2, G, GC, G^2C, G^3C	.528
9.	C and D	$C^2, G, GC, G^2C, G^3C, G^3C^3, D^2C, D^2GC$	.574
10.	C and G	$C^2, G, GC, G^2C, G^3C, D^2, D^2C$	.571
		Pre-Reading Completion Scores	
11.	--	C, C^2	.244
12.	C	D^3, DC^2	.296
13.	C	C, C^2, G	.246
14.	C and D	D^3, G^3D, DC^2	.309
15.	C and G	G, D^3, DG^2C^2	.307
		Post-Reading Completion Scores	
16.	--	C, C^2	.522
17.	C	C, C^2, D, D^3C^2	.557
18.	C	C, C^2, G, G^3	.561
19.	C and D	$C, C^2, G, G^3, D, D^3C^2, G^3D, GD^2$	.618
20.	C and G	$C, C^2, G, G^3, D, D^3C^2, G^3D$	.616

The second regression analysis at this stage consisted in going through the same procedure but this time using terms involving the grade variable in place of the difficulty variable. At the point where the cloze terms had been forced into the equation, the partial correlations involving the grade variable terms ranged from .20 to .35 and their corresponding *F* values ranged from 55.22 to 183.79. Line 3 shows the results when the analysis was allowed to run to completion. It can be seen that grade had a large effect on the regression, increasing the *R* from .15 to .39 and increasing the proportion of variance accounted for from 2.4 to 15.6 percent. Grade had a constant effect on the ratings as shown by the fact that the G^2 and G^3 terms entered the equation. Moreover, an interaction term, GC^2, entered the equation. The magnitude of its effects may be judged by the fact that it exhibited a correlation of .08 with the ratings when the effects of the other variables in the final equation were partialed out. Hence, the regressions at different grade levels differ not only in their heights but also they differ systematically in slope and possibly in shape.

The object of the third phase of the analysis was to determine if one of the variables, grade or difficulty, affected the regression of ratings on the cloze scores once the other two variables, difficulty and/or grade, had been entered into the equation. The first of these two analyses began with the terms shown in Line 2 of Table 3, forcing those terms into the equation to obtain the same equation that resulted from the second phase of the analysis. All of the terms involving grade, including the three-way interactions, were included in the matrix for this analysis. The partial correlations of these terms involving grade were then examined. They ranged from .10 to .33 and their *F* values ranged from 14.66 to 164.61.

When this regression was allowed to proceed until all, and only all, significant terms had entered, the equation shown in Line 4 was obtained. At first glance the terms in this equation appear somewhat surprising since the equation completely excludes the cloze terms. However, this is in keeping with the finding of the second phase that cloze and difficulty have highly redundant effects in this regression. This particular outcome was caused by a peculiarity of the design of the study that forced a zero correlation between grade and difficulty, while permitting cloze to have correlations of -.67 with difficulty level and .27 with grade. As a result, when grade entered it forced down the partial correlations of the cloze terms, leaving those of the difficulty terms unaffected. And when a term containing difficulty entered, the partial correlations of the cloze terms were depressed so far that they were deleted from the equation altogether and replaced by difficulty terms. The fact that cloze and difficulty are highly redundant in this regression can be further seen in the fact that if the cloze terms in Line 3 were replaced by difficulty terms, the resulting equation would be very similar to the equation represented in Line 4. However, it should be noted that this analysis did show that the addition of grade to the equation after difficulty

and cloze had been taken into account did affect the regression. And, again, adding the grade variable introduced an interaction apparently of about the same form as that observed in the second phase.

The second regression in this third phase consisted in forcing in the terms shown in Line 3 of Table 3 and then examining the partial correlations of the terms containing the difficulty variable. These correlations ranged from .01 to .09 and their corresponding *F* values from .07 to 11.79. When this analysis was run to completion, the equation shown in Line 5 was obtained.

From these analyses, it seemed clear that the difficulty variable had very little effect on the regression of the ratings on the cloze scores. The effect that it did have was merely to add a constant amount to the regressions. If the regression of the ratings on the cloze scores were plotted separately for each difficulty level, the result would be a series of parallel curves separated by fairly small intervals. Thus difficulty would have no effect on the level of cloze score at which the performance criterion is located by the model. Absolute heights of the curves do not influence the model. The grade variable, on the other hand, not only adds a constant to the regressions at each grade level but causes the regressions, themselves, to differ in shape at each grade level. Consequently, when willingness-to-study ratings are taken into account in setting a performance criterion, it is essential to calculate different performance criterion scores for students at each grade level.

The regression of the ratings on the cloze scores is shown in Figure 3. The equation on which these curves are based included only the cloze and the grade variable terms as independent variables. The equation was fit using both statistical and visual criteria. Terms were entered into the equation in an order determined by the relative sizes of their partial correlations with the criterion, and they were discarded from the equation if their *F* values dropped below .005. At an early stage of the analysis, the decision of when to stop admitting variables was made on the visual basis of what equation seemed to yield the best fit to the column means plotted at each grade level. These column means were obtained by ranking the students in each grade according to the sizes of their cloze scores and then calculating the mean and standard error of the mean ratings by the students falling into each 5-point interval on the cloze scale. No data were taken into account from the extreme portions of the range, the cut-off being placed at the point where less than .5 percent of the students in a grade level had a more extreme score. The object of the fitting operation was to keep introducing terms until the regression lines passed within one, or at most two, standard errors of each column mean and yet did not show localized dips due to column means that fell considerably above or below the means adjacent to it.

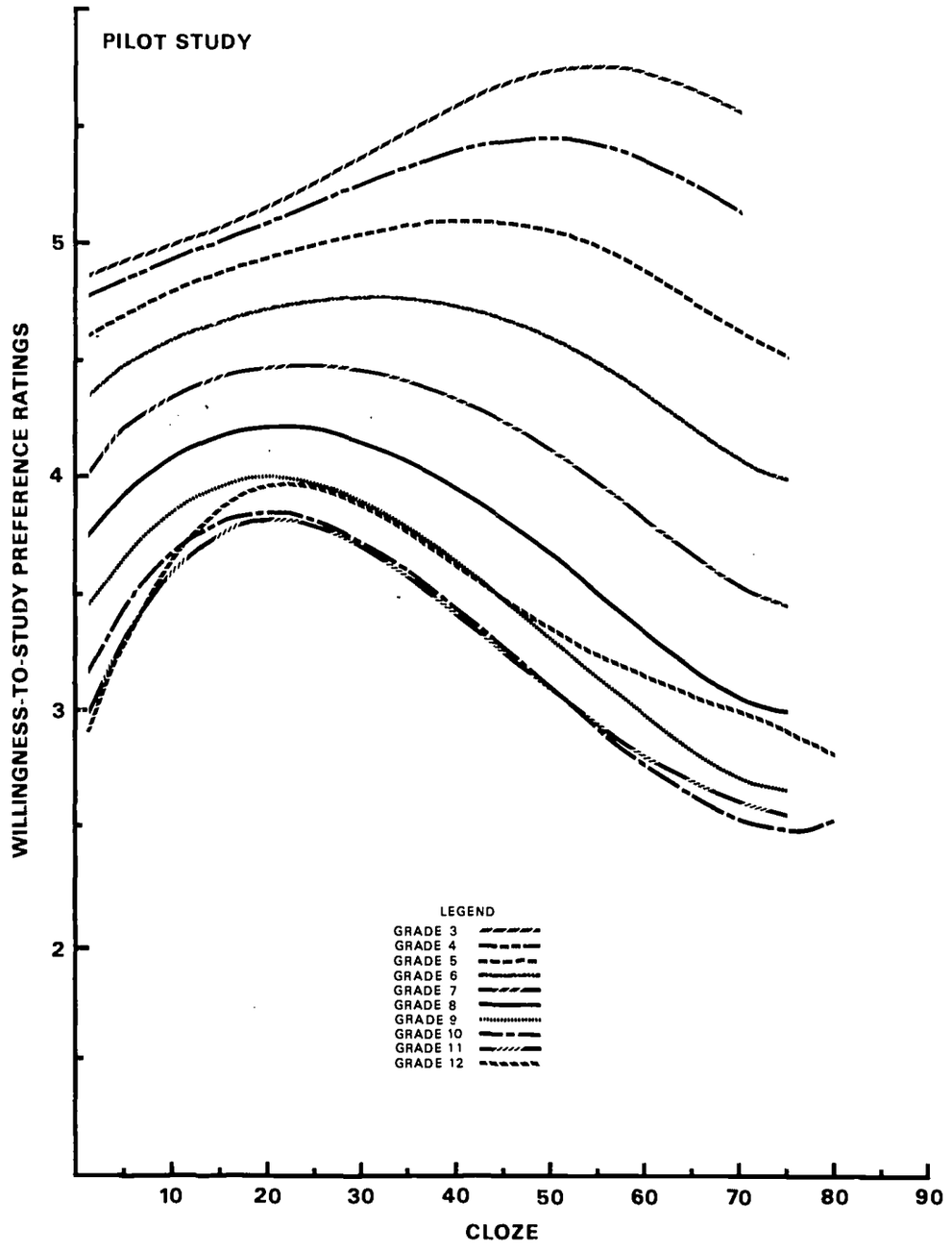


FIGURE 3. Regression of willingness-to-study preference ratings on cloze scores. Each curve represents the regression for students in one grade level. The vertical separations between the curves show that students at higher grade levels generally rated passages lower than the students in lower grade levels, regardless of the score they made on the cloze tests. The nonparallel nature of these curves indicates that different passage performance criterion scores are appropriate for students at different grade levels.

Experience with this rather laborious procedure showed that, at least when large numbers of students are available, the procedure may be unnecessary. On the one hand, merely having large numbers of students did not greatly improve the fit of the regression equations when the terms of the equation were entered merely on the basis of their statistical significance. The problem is that present regression analysis methods weight columns primarily according to the number of students having cloze scores of that value. Hence if there were identical numbers of students obtaining each cloze score, the curve obtained would represent each level of cloze performance fairly well. However, when the scores fall into a distribution that appears to be somewhat normally distributed, as in these data, the curve will fit well only around the cloze mean, as was seen in Figure 1. On the other hand, inserting additional terms well past the point where the terms seemed to account for significant amounts of variance failed to produce the anticipated effect of forcing many localized dips into the regression curve. Localized dips did occur, but only beyond the scores defined as the observed range, that is, beyond the minimum and maximum scores that enclosed 49.5 percent of the subjects on either side of the mean within a grade level. Somewhat beyond that range, however, the curves exhibited highly irregular forms. The equations plotted in the figures to be presented, then, were finally obtained simply by setting the parameters of the regression program to enter all, and only all, terms having an *F* value of .01 to enter and .005 to be deleted from the equation. The resulting equations are presented in Table 4.

Figure 3 shows the ratings for each grade level plotted against cloze scores. Each curve followed roughly the shape predicted by the Dember-Earl theory. That is, the students rated passages at an intermediate range of cloze score higher than they rated the materials at somewhat higher and lower cloze scores. However, this was only one of the strong systematic effects observed in the data, and so it seemed clear that the theory will have to undergo some elaboration in order to explain preference for instructional materials. The theory provided no grounds for anticipating (a) that older students would generally rate passages at every level of cloze score lower than the younger students rated them; (b) that some of the curves would be relatively flat and plateau-like rather than sharply peaked; and (c) that the curves would peak at progressively lower levels of cloze performance as the students got older, that is that the peaks would shift to the left on the graph.

Each phenomenon is subject to a number of alternative explanations, some of which will be briefly described here. Although the purpose of the present studies was not to get involved in building motivation theory, the unanticipated results demanded theoretical interpretation since the various interpretations that can be placed upon these results had different implications for the use of preference ratings in identifying performance criteria. This discussion will be a summary, and a more detailed discussion made the topic of a separate report.

TABLE 4

Equations Plotted in the Figures

Willingness-To-Study Ratings = $.2981G - .05999G^2 + .0001620G^4 + 26.216C^3$
$- 36.250C^4 + 14.356C^7 + .4679CG + .003655CG^3 - .3743C^2G^2 + .01161C^2G^3$
$- 1.3513C^3G + .5444C^4G^2 - .04253C^6G^3 + .01660C^7G^3 + 4.4826$

Information Gain Scores = $.02152G + .7317C + .2630G^2 - .00001611G^4$
$- 3.5799C^2 + 1.8278C^3 - 2.6359C^5 + 4.3149C^8 + .07313CG - .0003107CG^3$
$+ 1.0509C^2G - .1794C^2G^2 + .007429C^2G^3 + .3992C^3G - 1.1960C^6G$
$+ .1052C^6G^2 - .003191C^6G^3 - .02546$

Pre-Reading Completion Scores = $.001947G - .2924C + .00006203G^2 + 1.4941G^5$
$+ .2236CG - .03463CG^2 + .001390CG^3 - .06350C^2G + .01762C^2G^2$
$- .02691C^3G - .001511C^3G^3 - .6189C^5G + .06094C^5G^2 + .02655C^7G^2$
$- .003180C^7G^3 + .007643$

Post-Reading Completion Scores = $.04803G + .9423C - .000008682G^4$
$- 4.4079C^2 + 2.8183C^4 + 2.3098C^8 - .0001405CG^3 + 1.8020C^2G - .2316C^2G^2$
$+ .008795C^2G^3 - 2.0293C^5G + .2091C^5G^2 - .009390C^6G^3 + .04140C^7G^2$
$- .06392$

Consider the phenomenon that students at higher grade levels generally rated all materials lower than the students did at lower grade levels. While the regression of lower grade ratings ranged from neutral, a rating of 4, to mild enthusiasm, a rating of 5, the higher grade students' regressions ranged from dislike, a rating of 2, to neutral. The first explanation of this effect regards this result as an index of a general tendency for students to learn to dislike instructional materials and school in general as they proceed through the grades. That is, along with whatever else the schooling process is thought to teach, it also teaches the student to dislike that process itself. This type of interpretation finds some support in studies such as that by Cornell, Lindvall, and Sciupe (1953), in which they observed that the reactions of students to teachers and teachers to students became increasingly negative over the range of grades 4 through 10.

A second interpretation might be that a student's interests grow increasingly specialized with age and that he identifies fewer passages that deal with topics still of high interest to him as he gets older. Thus, students at all grade levels maintain a uniformly high level of enthusiasm for studying something, but the range of things the student is enthusiastic about narrows systematically with age. This theory finds some support in the common observation that people's interests become increasingly specialized with age, and it also accounts for the fact that there were a substantial number of student ratings above 4 at every grade level.

A third possible explanation is that the rating task, itself, systematically changed in meaning. Whereas younger children are generally observed to express enthusiasm for adult-defined tasks and their peer group regards this as socially acceptable behavior, older students, on the other hand, tend to react negatively to such tasks and sometimes exercise social sanctions against those of their peers who react positively--the charges of being an apple-polisher or a greasy-grind being matters that cannot be taken lightly since they carry the threat of some form of social ostracism.

None of these interpretations can be excluded on the basis of the evidence presented in this or other studies known to the author. However, these interpretations are theoretically rococo since they involve explanatory devices that do not express the response as a function of properties of the stimulus materials and the preference rating task. Consequently, while such explanations may be of interest for theories used to make policy decisions in social engineering, they are of secondary interest for making policy decisions for engineering instructional materials. If for no reason other than the aesthetics of obtaining parsimony of theory, it seems desirable, then, to pursue here only explanations that relate the response levels to dimensions of the materials. Moreover, it appears at least possible that only the latter type of explanation may be required to account for most of the observed results.

The fourth interpretation, the one preferred here, is based on a further analysis of the Dember-Earl dimension of the information content of the stimulus. It is proposed that there are two kinds of information in the stimulus, structural information and topical or semantic information. The position taken here is, first, that the Dember-Earl theory provides a reasonably accurate description of the relationship between the amount of structural information in a stimulus and the indices of interest or preference. That is, the regression has the stylized general form of the absolute value equation

$$(31a) \quad I = i_c - \alpha \mid S - C \mid$$

or the quadratic form that is more probable in actual observations,

$$(31b) \quad I = i_c - \alpha (S - C)^2 \quad ,$$

where I stands for some measured index of interest such as preference ratings or inspection time, C for the subject's complexity level measured in some metric, S for the structural complexity of the stimulus measured in the same metric as C, i_c for the height of the curve at point C, and α is a constant of curvature or slope. Second, this position asserts that interest decreases as a function of a person's familiarity with the topical or semantic content of the stimulus. That is,

$$(32) \quad I = [i_c - \alpha (S - C)^2] - \beta T + i_T \quad ,$$

where the brackets enclose the right member of equation (31b), T is some measure of the subject's familiarity with or knowledge of the topical or semantic content of the stimulus, β is a slope constant, and i_T is the intercept of the plot of I on T. Although this expression shows the regression of I on T as linear, this is merely for the sake of temporary simplicity, for it seems more likely that it, too, would follow a quadratic or even a cubic function of some sort. Within the range of data represented in the present study, however, only a quadratic function seems justifiable.

Two initial points must be made to justify this extension of the Dember-Earl model. Structural complexity and topical familiarity are both operationally definable and independently manipulable. A subject's familiarity with a stimulus can be manipulated while holding constant the stimulus complexity and varying his familiarity with the stimulus or vice versa. By repeatedly exposing the subject to a stimulus or to various elements or other transformations of the stimulus, his familiarity with its content can be altered without changing the physical characteristics of the stimulus. Thus, for example, a person's familiarity with the semantic content of a passage can be altered by having him read all or part of it some number of times or by exposing him to formally different materials that provide explanations of the same topic, but use other language in doing so. Topical familiarity, then, is measurable either in terms of the amount of prior

exposure of a person to the stimulus or to stimuli that present the same information through some sort of transformed version of the stimulus, and it may also be measured in terms of the amount of knowledge he can exhibit on that topic prior to his exposure to the stimulus.

On the other hand, structural complexity, as this term is used here, refers to the manner in which content is presented. That is, there are many alternative forms in which essentially the same topical or semantic content can be expressed in a given mode of communication; each of these alternative forms can normally be expected to transmit that content with differing degrees of effectiveness; and systematic structural relationships among the alternative forms can be related to the degrees of effectiveness with which the content of the alternative forms is learned. For example, the sentences

(33) The boy climbed on his horse.

(34) The steed was mounted by the boy who owned it.

seem to transmit nearly the same content but probably differ in the effectiveness with which they do so, the degree of effectiveness being measurable by various types of tests. Readability theory provides a number of complexity measures that can predict much of this variation in effectiveness. Thus, topical familiarity and structural complexity each represent an operationally definable dimension of a stimulus, and each is manipulable independently of the other.

To return to the interpretation of the phenomena observed in this study, variations in the curve for each grade level were regarded as representing, to some degree, the effects of structural complexity on the preference ratings. Thus, they exhibited some similarity in shape to an inverted V. The large vertical separations between those curves, on the other hand, were interpreted as attributable primarily to variations in topical familiarity among the students at various grade levels. The younger students were probably less familiar with the topics discussed in the passages used in this study than the older students, and therefore rated the passages relatively higher than the older students did. Had the topics of the passages been equally unfamiliar to students at all grade levels, the curves would have been identical.

The second phenomenon noted in Figure 3 was that the curves of some of the grade level groups were relatively flat and plateau-like. This is attributed both to the theory and to the fact that the cloze procedure provides a measure that confounds the effects of topical familiarity and structural complexity. That is, a student's scores on cloze tests vary both with his knowledge of the topics of the passages and with the degrees of complexity of the language in the passages. One effect of this would be to flatten out the curves observed here. Two students within a given grade level might have different scores on a cloze test over a passage. But because those scores could represent varying combinations

of familiarity with the topic and ability to cope with the structural complexity of that passage, their preference ratings could be identical, thereby causing the curve to flatten. Had a measure of topical familiarity more sensitive than grade level been used, presumably the curves might have peaked somewhat more sharply. Another effect that could flatten the curves would occur if students were individually responding to different dimensions of the passage--some to writing style and others to difficulty, for example. There is no necessary reason why all types of responses, when regressed on cloze, should all peak at the same point. Consequently, the effect would be to flatten the curves. It should also be mentioned again that Kamman (1966) found that students' preferences for easy materials correlated negatively with a measure of debilitating anxiety and positively with a measure of facilitating anxiety. Effects of this sort would also tend to flatten the curves; however Kamman's data also showed that the anxiety ratings showed very similar correlations with the students' cloze scores, making it uncertain whether anxiety produces an effect on these curves that cannot be attributed merely to the correlation of cloze and anxiety scores.

The third phenomenon noted in Figure 3 was that the curves of lower grade students peak at systematically higher levels of cloze performance than the curves of the older students. This may be attributable, in part, to the manner in which this study was instrumented. The passages used exhibited some degree of correlation between the structural complexity of the language in the passages and the likely familiarity of the student with a passage. The passages at the easiest levels, for example, were written in language of the type variously labeled as primerese or Dick and Jane English, and they dealt with topics such as young children playing at games. The passages at the more difficult levels, however, contained highly complex grammatical structures and technical terms and dealt with topics such as the crystal structure of rock.

Thus, most students regardless of grade level tended to make low cloze scores on passages that were both low in familiarity and high in structural complexity. This seems to have caused the curves for all grade levels to lie closest together in the region of low cloze scores. However, as higher levels of cloze performance were reached, these cloze scores were obtained on passages that remained relatively unfamiliar for the students in the lower grades and simultaneously the passages approached the students' complexity levels, and so the trends of their curves remained generally upward at these somewhat higher cloze scores. When these cloze score levels were reached by the students in upper grades, however, those cloze scores tended to be obtained on passages that dealt with topics that were highly familiar to the students. Moreover, the topical familiarity effect seems to be the stronger of the two effects, since the variability of the regression lines between grades is much greater than the variability of the regression line for a single grade. Thus, the students in the higher grade levels who made fairly high cloze scores exhibited extremely low ratings even though the passages may have been at their complexity level.

This confounding in the passages, then, seems to have been responsible for the fanning out of the curves or the shifting of the peaks, depending on how it is viewed.

There was one other strong effect in this set of curves--the tendency for the decline between grade levels to level off and begin to reverse at the high school level. This effect suggests that preference ratings are not linearly related to topic familiarity. Since somewhat the same effect also appeared in the next study in this series, this can probably be regarded as a genuine effect, and the model shown in (32) should properly be modified by including at least a quadratic term for topic familiarity. However, this conclusion could not be drawn from the present data, due to the fact that students in the high school were drawn from a different and less able population than the students in the lower grades.

The results of this study of preference ratings had four implications for the performance criterion model. First, it was clear that the curves for each grade level had different slopes and shapes. Thus, it is essential to calculate a different performance criterion for each grade level. Second, the facts that the regression curves were relatively flatter than theory predicted and that two, rather than one, dimensions of the passages were required in order to interpret these preference ratings made it seem advisable to explore the matter further to determine if students could distinguish among still other dimensions of the passages in their ratings. Third, in a sense these results validated the cloze test to serve as the basic measure for a passage performance criterion since it is apparently a confounded measure of both of the passage dimensions so far identified as influencing preference ratings. Finally, because the difficulty variable was highly redundant with the cloze variable, it seemed unnecessary, at least with respect to the willingness-to-study variable, to calculate separate criterion scores for different levels of passage difficulty.

Information Gain Scores: The regression of information gain scores on the cloze scores was submitted to a similar set of analyses. Line 6 of Table 3 shows the results obtained when just the cloze scores were permitted to enter the regression equation. The multiple correlation of .48 seemed fairly high in view of the fact that both the cloze and the completion tests contained few items. When this initial equation was used in the second phase of the analysis to determine the effects contributed by difficulty, the terms derived from difficulty exhibited partial correlations ranging from .14 to .19 and corresponding F values from 29.20 to 49.65. However, it may be seen from Line 7 that only the D^2C term entered and that the increase in the multiple R was fairly small, accounting for only about 3 percent more of the variance in information gain. When the terms derived from grade were analyzed along with the initial equation, the partial correlations ranged from .01 to .30, with corresponding F values of .19 to 126.80. When the analysis was

permitted to run to completion, the grade terms increased the multiple R by a somewhat greater amount than the difficulty terms had; they increased the proportion of variance accounted for in the information gain scores by roughly 10 percent.

The second stage of this analysis showed that grade contributed a considerable amount to the regression of information gain on the cloze scores. Moreover, grade interacted with cloze in these regressions, indicating that the curves plotted for each grade level would exhibit different slopes and possibly different shapes. The effects of difficulty on this regression, however, were problematic. Apparently the difficulty terms are highly redundant with cloze scores in this regression, also, since they added only a negligible amount to the regression. And this redundancy makes it difficult to interpret how the D^2C interaction may have functioned in this analysis. On the one hand, difficulty may have represented a measure of something that was substantively different from the processes measured by the cloze scores. However, this is cast in doubt by the fact that, when the cloze scores were held constant, the partial correlations between gain and the difficulty terms were fairly low, and only the one difficulty term entered the equation when the analysis was allowed to run to completion. On the other hand, it seems possible that the difficulty variable merely represents some sort of transformation on the cloze scores, a transformation that may be somewhat more complex than the integer powers of cloze and mathematically less correlated with them. This possibility gains some support from the outcome of the analysis of the ratings of willingness-to-study and even stronger support from the fact that passage difficulty is defined as the mean of the cloze scores over a passage. Consequently, at this stage of the analysis it seemed doubtful that difficulty terms affected the regression in any way that cannot also be obtained by using transformations of the cloze scores in the analyses.

The value of the third phase of this analysis is cast in doubt because of the problematic status of the difficulty terms. When the equation shown in Line 7 was computed in a matrix that included just the terms from Line 7 and the terms that involved grade, the partial correlations of the terms containing grade ranged from .03 to .23 with corresponding F values of 1.37 to 71.12, and the equation that resulted from permitting those terms to enter had an R of .57, as is shown in Line 9. When the terms in the equation shown in Line 8 were analyzed in the context of a matrix containing the terms derived from difficulty, the partial correlations ranged from .13 to .25 with corresponding F values of 85.11. An equation containing the terms shown in Line 10 resulted from the completion of this analysis. It should be noted that the terms in Lines 9 and 10 were highly similar in spite of the fact that they began by forcing in different sets of variables.

This third phase of the analysis showed that difficulty and grade each had some effect on the regression of gain on cloze scores. However,

it shed little light on the problematic nature of the difficulty terms. It is true that the difficulty terms did not increase the multiple *R* by quite as much as the grade terms; however, it is doubtful that much significance can be assigned to the difference between an *R* of .571 and one of .574. Perhaps it is significant, though, that there was only a 4 percent difference in the amounts of variance accounted for by regressions that entered all, and only all, significant cloze and grade terms and a similarly calculated equation that permitted difficulty terms to enter, also. Consequently, in view of the relatively small effects of difficulty on the regression of gain on cloze scores and the problematic status of difficulty as a measure of something that differs qualitatively from what is measured by the cloze scores, it did not seem advisable to employ difficulty further as a variable along which performance criteria might need to be differentiated. Grade level, on the other hand, had a considerable effect on the regressions of gain on cloze. It showed seemingly strong interactions with the cloze variable, suggesting that the regression curves differed systematically across grades. Thus, these analyses made it appear desirable to elaborate the criterion selection model to identify a separate performance criterion for each grade level of student.

Figure 4 shows a graph of the equation finally fit to the regression of gain on cloze and grade. The regression equation itself is shown in Table 4. The fitting was done using the same combination of statistical and visual criteria used to fit the equation graphed in Figure 3. This set of curves exhibit three characteristics of major interest: (a) the curves for the different grade levels differed in shape and slope; (b) there was a large systematic effect of grade level on the heights of the intercepts; and (c) there was no region of zero slope in the lower range of cloze scores.

The fact that the curves differed in shape and slope for the different grade levels seemed to dictate again, as in the regression of the ratings on cloze, that the criterion selection model be elaborated to obtain a different performance criterion for each grade level. However, peculiarities in the shapes of these curves suggested that the matter be examined more closely. First, the fact that the intercepts were fairly high, for example, contradicted the results from the earlier field trials that had shown intercepts for just the pre-reading scores near zero for students at grades 4 and 10. Second, it had shown a short region of near-zero slopes in the curves for students falling into the lowest range of cloze scores, much as the first study in this series had shown. And third, it had shown that taking the same completion test both before and after reading a passage increased the students' scores on the post-reading administration of the test. Thus, there was strong reason to suspect that this set of curves did not represent an accurate picture of the results that would be obtained when these carry-over effects were eliminated in a more elaborate testing design. However, the field trials had also shown that the curves of the two grade levels

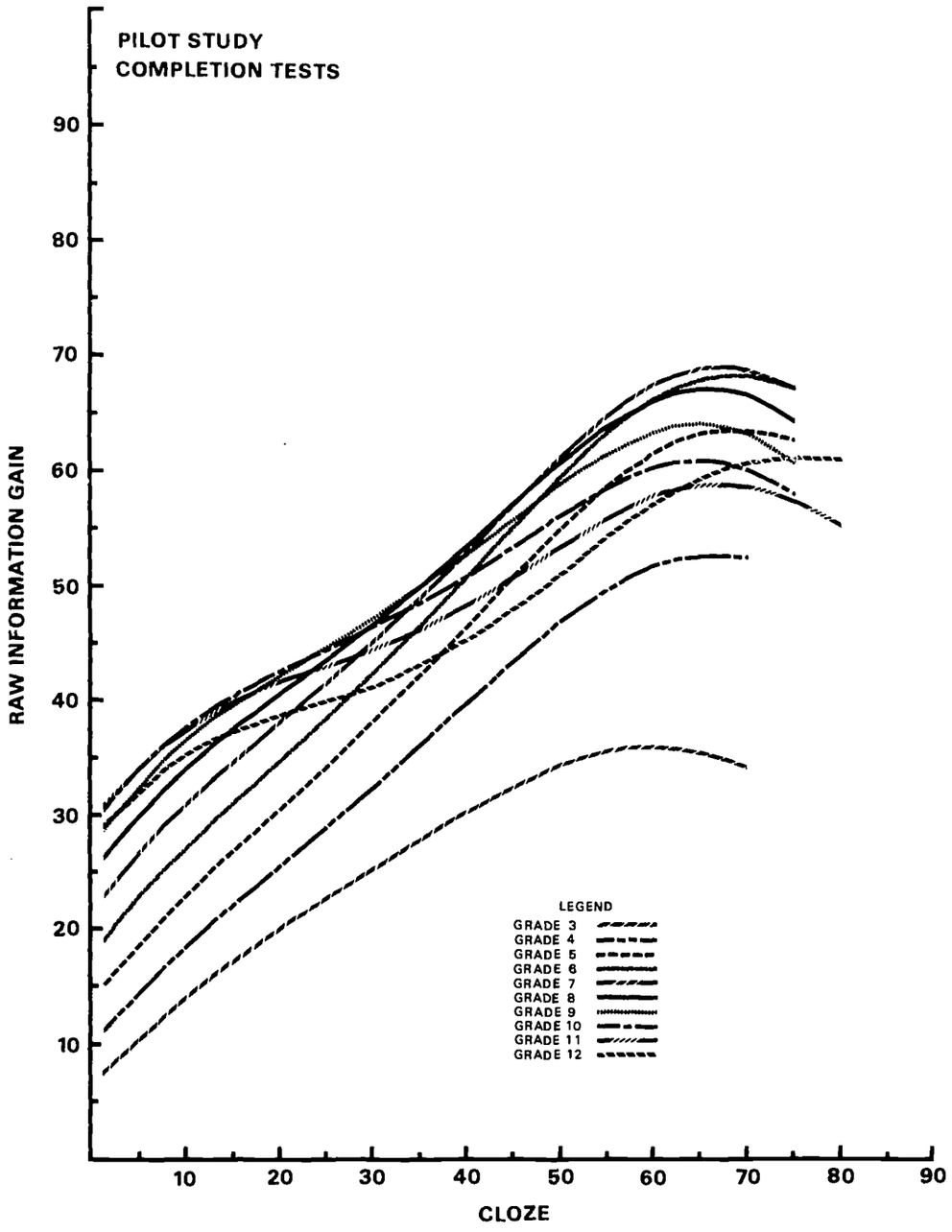


FIGURE 4. Regression of raw information gain scores on cloze scores. Each curve represents the regression for students in one grade level. The vertical separations between the curves show that students at higher grade levels generally made higher gain scores, regardless of the score they made on the cloze tests. The lower slopes of the curves for the high school students may have represented a lack of continuity in the population of students sampled.

tested, grades 4 and 10, seemed to diverge in a manner that suggested that students at higher grade levels would show higher levels of information gain curves, much as they did in the data being examined. Consequently, it was speculated that the carry-over effect influenced the curves primarily by increasing them in the lower ranges of cloze scores and that the curves would probably continue to differ in slope when the possibility of a carry-over effect was removed. However, this issue was no longer critical for the design of the criterion selection model since the outcome of the study of the regressions involving the preference ratings had already made it clear that future studies would have to be designed so that a different performance criterion could be calculated for each grade level.

These data provided indirect evidence that the regression identity assumption held to a reasonable degree for the information gain regressions. If the regression identity assumption held perfectly for tests of these types, the regression calculated for each pair of passages would have all lain along a single regression surface, and this would have resulted in a fairly high multiple correlation coefficient. On the other hand, violations of the assumption would have taken the form of the regressions exhibiting bivariate means lying well away from the regression surface, with the individual regressions exhibiting slopes that projected at angles to the regression surface. The effect of these events would be to lower the size of the multiple correlation coefficient. Thus, the coefficient observed, .57, may be taken as a measure of the degree to which the regression identity assumption held for these data. It must, however, be modified since a number of other effects were also involved. On the one hand, the size of this coefficient underestimates the degree of correlation that actually existed because the tests were all quite short and had low reliabilities, and because the passages within difficulty levels were not perfectly matched. On the other hand, the coefficient was probably somewhat inflated by the fact that the between-passage variation was added to the within-passage variation in these regressions, and also by the fact that the scores were obtained from students ranging widely in grade level. However, the latter effects were accounted for as systematic variation due to difficulty and student. Consequently, the regression identity assumption can be said to hold to a reasonable degree of approximation in these data, even though the assumption could not be tested directly.

Pre- and Post-Reading Tests: The scores on the pre-reading and post-reading tests were submitted to analyses of the same types as those used to analyze the preference ratings and the gain scores. The purpose was to provide checks on some of the interpretations made of these earlier regressions more than to obtain information directly relevant to the performance criterion selection model. Table 3 presents the results of the regression analyses, Table 4 presents the equations that were graphed, and Figures 5 and 6 show those graphs.

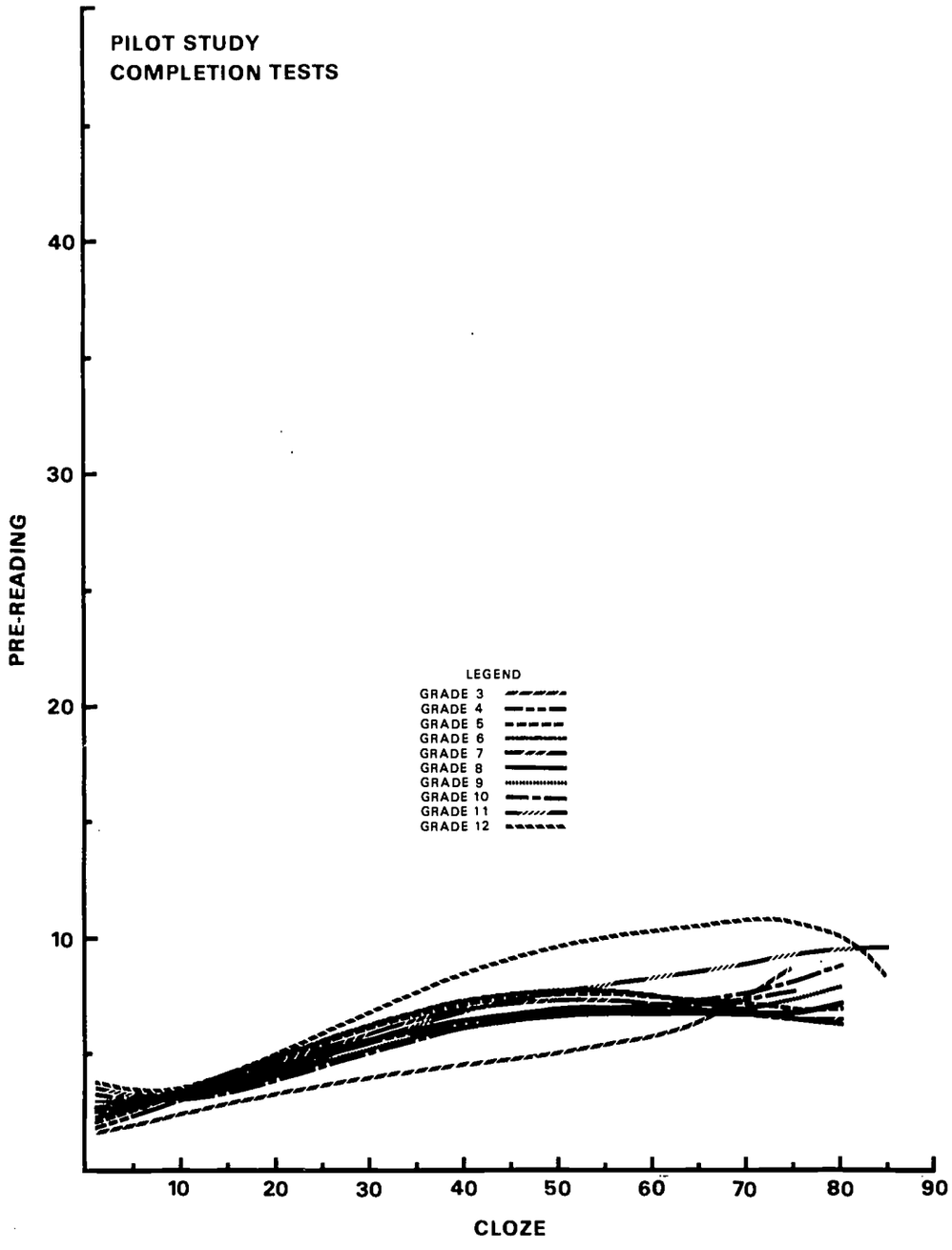


FIGURE 5. Regression of pre-reading completion test scores on cloze scores. Each curve represents the regression for the students in one grade level. The small vertical separations between the curves show that student grade level had a slight but systematic effect on the pre-reading scores.

The first point of interest in these results is that they provide support for the contention that the difficulty dimension on which the passages in this study were arrayed represented, to some extent, a measure of a student's likelihood of having knowledge of the content of the passage without having read it. The strongest correlation of the pre-reading tests was with the difficulty variable, as may be seen from the results in Lines 12 through 15 of Table 3. Since these tests were administered in a manner that measured the student's knowledge of the contents of a passage prior to his having read it, this set of regressions provides support for the contention that the passage difficulty dimension in this study represented, at least to some degree, a dimension that arrayed passages according to a student's probable familiarity with the topics dealt with by the passages.

However, these results provided evidence that the difficulty dimension also arrayed the passages according to their structural complexity. If it were assumed that much of the content familiarity component could be removed from the post-reading scores by subtracting from them the pre-reading scores, then the gain scores should provide a fairly good measure of the structural complexity of the passages. There is some evidence that this was actually the result. When the terms in the equations involving gain in Table 3 are compared to those involving the pre-reading and post-reading tests, it can be seen that the effect of deriving the gain scores was to subtract many of the difficulty terms from the post-reading regression to obtain the information gain regression. Thus, the gain variable could be said to represent, at least in part, a measure of structural complexity. Since the difficulty variable also entered some of these regressions, it seems possible that the difficulty variable also arrayed the passages according to their structural complexities.

Summary and Evaluation

This study was conducted to investigate some of the methodological and substantive problems that had to be dealt with en route to developing a criterion selection model. The methodological problems consisted chiefly of developing and applying better controls over the test-writing and test-scoring procedures and of developing an efficient data gathering design. The really fundamental methodological problem of devising a direct test of the regression identity assumption was not solved and remained unsolved at the time this report was being prepared. The substantive problems consisted of determining whether different performance criteria must be derived for students at each grade level and materials at each difficulty level. In the course of the study it was also necessary to make a digression into the study of motivation theory and then to reassess the methods used to measure preferences for materials. What follows is a brief evaluation of the main results of this study.

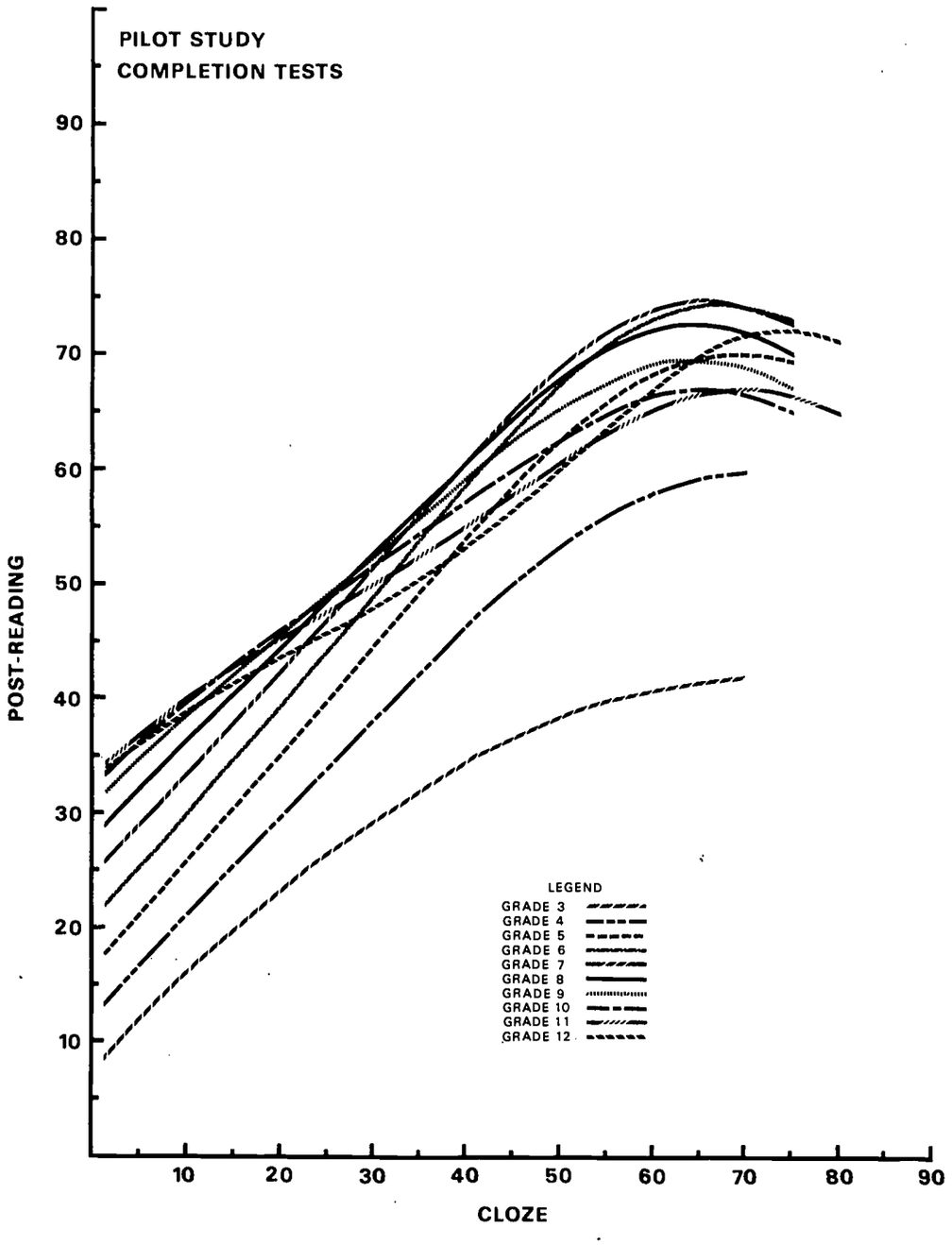


FIGURE 6. Regression of post-reading completion scores on cloze scores. Each curve represents the regression for students in one grade level. The vertical separations of the intercepts of the information gain scores may have been traceable to the similar separations observed in this regression.

Item-Writing Procedures: The item-writing procedure employed was a sequential procedure that consisted in diagramming the syntactic structures of the paragraphs, numbering the nodes from which questions could be derived, randomly selecting the first nodes to be tested, deriving the questions, identifying further questions that could be derived from elements of the question stems, randomly selecting those stems, and deriving the questions. This procedure should be evaluated against the criterion of whether it produces tests that are operationally replicable. Test-writing procedures that fail to meet this criterion can only produce results that have no more status than that of an unverifiable personal opinion in making public policies such as those represented by performance criterion scores.

The procedures used fell somewhat short of meeting this criterion. The first problem arose because it was not always possible to mark questions with single wh- pro words. As a result, the test writer had to employ some personal judgment in wording the wh- phrase. Second, the process of deciding when further elements of the question stem can be deleted was not rigidly governed by workable rules; consequently the test writer had to exercise some personal judgment on this matter. However, the largest proportion of the questions, roughly 90 percent, were completely replicable, and the writing of the remainder was governed by only somewhat looser rules. Consequently, the test-writing procedures used in this study were regarded as producing results that would be replicable within fairly close tolerances.

Item-Scoring Procedures: The scoring procedure was a method of classifying responses as being either derivable or not derivable from the passage. Those that were derivable are counted correct and those not derivable were scored wrong. This procedure produced highly replicable results. Of the more than 26,000 responses possible in this study, a total of less than a dozen responses were identified as presenting a problem for this scoring theory.

Differentiation of Performance Criteria: The results clearly indicated that it is necessary to calculate separate performance criterion scores for students at every grade level. A strong interaction occurred between cloze and grade when the willingness-to-study ratings served as the dependent variable. A similar interaction may have occurred also when information gain served as the dependent variable. However, this evidence came from a small ancillary study rather than from the main study. It seemed likely that the interaction may have been made ambiguous in the main study by the fact that there seemed to be a large carry-over effect from administering the same test as both a pre-reading and post-reading test.

It appeared unnecessary to calculate different criterion scores for different levels of passage difficulty. Although passage difficulty contributed a significant amount of variance to the regression of gain

and preferences on cloze, these were relatively small contributions. Moreover, since difficulty is defined as the mean of cloze scores on a passage, it is difficult to separate the cloze and difficulty measures conceptually. Consequently, it appears that little would be gained by calculating different criterion scores for different levels of passage difficulty.

Measurement of Interest: The results of the regressions of preference ratings on the cloze scores showed that it was necessary to postulate at least two dimensions of the passages as influencing the ratings --the familiarity of topics of the passages and the structural complexity of the passages. As a result, it seemed advisable in subsequent studies to develop additional rating scales to attempt to differentiate students' responses to these two types of stimuli.

Regression Identity Assumption: The regression identity assumption received some further support even though it was not directly challenged by a definitive test of its validity. The preference ratings and the completion test scores exhibited correlations with the cloze scores that could not be accounted for by the effects of differences in passage difficulty or differences in student grade level. Moreover, the regression curves plotted were obtained by pooling the scores from many tests, and yet the curves that resulted did not greatly differ in shape from the ones obtained in the first study. Consequently, there seemed to be adequate grounds for proceeding on the assumption that the regression identity assumption was satisfied to a reasonable degree of approximation by the tests employed in this study.

STUDY III

Introduction

This study attempted to obtain regressions that would be useful for at least tentatively identifying performance criterion scores. Prior to the completion of the preceding study, the feasibility of establishing a performance criterion seemed too doubtful to risk many resources on attempting either to systematically analyze and model the concept of a performance criterion or to gather data that could be used directly in identifying a criterion score. It was clear that the feasibility depended upon the validity of the regression identity assumption, particularly for the comprehension tests. It was also clear that the comprehension tests had to exhibit instructional conformity and behavioral consistency as prior conditions for satisfying the regression identity assumption. While comprehension tests could be constructed by methods that permitted those conditions to be approximated, it was uncertain that even meeting these prior conditions perfectly would be sufficient to produce regression identities. The preceding study, however, provided evidence that the regression identity assumption is probably met by tests of the types used in these studies.

Objectives: Consequently, specific funding was sought and studies were commenced to achieve three objectives. The first was to analyze and model the concept of a rationally derived performance criterion. Most of the results of that work, to date, were described in preceding sections of this report. The second objective was to further analyze the preference ratings of passages and to obtain the regressions of those ratings and of information gain scores on cloze scores. The results of these efforts furnish the principal content of this section of the report. The third objective was to further develop the procedures for constructing comprehension tests. Some progress was made on this objective, but it did not happen in time to incorporate the improved techniques into the instrumentation of these studies. So, while the information gain regressions in this study are probably replicable enough to obtain fairly useful performance criterion scores, their replicability cannot be certified with all of the operational rigor that is desirable in an operation that serves as evidence in the making of policy.

Analysis of the Preference Ratings: Further efforts to analyze preference responses to passages began by interviewing students about their reasons for giving passages the ratings they did and reviewing some of the relevant theory in the field of rhetoric. When a number of students of different ages were questioned about their preference ratings of passages, the responses seemed to fall into three categories. The first category dealt with the student's like or dislike for the subject matter of the passage. This category contained responses such as *This paragraph is about chemistry and I don't like that course* . The second

category of statements dealt with their reactions to the style, such as *It talks down to you* or *It's too stuffy*. The third category of remarks dealt with the grade level at which the materials were taught and included responses such as *That's too young for me* or *We will get to study this next year*. The valuative statements that accompanied the remarks in this category generally suggested that the most desired passages contained materials that the student anticipated studying at the grade levels above his own and that the least desired were those dealing with topics taught at much lower grade levels. It seemed possible to analyze the responses to style into a considerably greater number of seemingly independent categories, formality-informality and inductive-deductive dimensions, for example. However, it could not be determined that all of the students in the age range of these studies had developed concepts corresponding to these dimensions of prose nor did there seem to be clear consistency to the values the students placed upon these dimensions.

The rhetoricians segment their subject matter along conceptual lines roughly paralleling these response categories (see Martin and Ohmann, 1965, for example). In broad terms, they address themselves to dealing with the prior preferences and biases of an audience, with the aesthetic stylistics of the discourse, and with setting the conceptual level of the discourse for audiences with differing amounts of prior knowledge on a topic. Each of these is regarded in rhetorical theory as being an independently manipulable dimension of the discourse and each is regarded as having a systematic effect on the attitudes of a reader toward the materials.

These two lines of evidence suggested that at least three scales should be added to the willingness-to-study scale used in the preceding study. These were scales to measure preference for the subject matter of the passage, preference for the style of the passage, and preference for the maturity or difficulty level of the passage.

The study by Kamman (1966) had provided evidence that the criterion selection model should also be elaborated to take into account the different uses made of materials in instruction. He showed that a subject's patterns of preference ratings of poems differed markedly, depending on the use for which the poem was being evaluated. Consequently, it seemed necessary in the present study to obtain ratings of the passages when they were being considered for use as a textbook, as a reference, and as a book read voluntarily, such as a book that a student might check out of a library and read for his own pleasure. Only the scales measuring preferences for style, difficulty, and willingness-to-study were employed to assess preferences under each of these three use conditions.

Test Construction Procedures: The test construction procedures used in the present study represent little improvement over those employed in the preceding study. A considerable amount of effort was expended in an effort to solve some of the test-writing problems encountered in the preceding study. However, the result of this work was to show that the primary

source of the difficulty lay in a deficiency in transformational theory of grammar, the deficiency being that the deep structure of sentences had not been adequately defined. One effect on the test-writing procedures, for example, was that it was impossible to determine how far the constituent deletion process can be carried in questions because the current transformational-generative grammars did not provide an adequate definition of what constitutes a minimal sentence or question. It now appears that the solutions to many of these problems can be developed from the case grammars proposed by Fillmore (1968). However, the efforts to achieve this solution are still in the formative stage.

Procedure

The testing design used in the preceding study proved to be highly efficient and so was retained for this study, but with three major alterations. First, each of the four types of tests taken by a student was made from a different passage. This was done in order to eliminate carry-over effects that resulted from administering the same test as both a pre-reading and a post-reading test within the short time spans involved in these testing operations. The second alteration was to employ passages that could be said to be more representative of instructional materials. The third alteration was to employ longer passages and comprehension tests in order to obtain more reliable results on the cloze and the information gain tests.

Passages: Eight sets of passages representing eight levels of difficulty were used in this study. There were four passages in each set matched fairly closely for cloze difficulty. These passages were selected from a sample of 330 passages whose relative cloze difficulties had been determined in a different study (Bormuth, 1969a). The 330 passages had been drawn from instructional materials used in schools using a sampling grid for a stratified random selection procedure. The grid consisted of a subject matter factor having ten categories of content and a level-of-school-usage factor having five levels, grades 1-3, 4-6, 7-9, 10-12, and college level. Each passage contained approximately 110 words and each represented the introduction to a randomly selected section of a textbook or some other form of instructional materials.

The passages were drawn from this sample by ranking the 330 passages in the order of their cloze difficulties, marking the median points of the four passages at either extreme of the distribution, dividing the range intervening between these two medians at six equally spaced points, and selecting the four passages closest to each of these eight marks. Three restrictions were placed on drawing a passage; the first was that no two passages from the same content category could be drawn from the same or adjacent difficulty levels. The second was that no more than four passages in the entire set of 32 could come from the same content category. The object of these restrictions was to reduce effects that

might arise from a correlation of content category with difficulty level. The third restriction was that the passage drawn be part of a continuous discourse of at least 250 words in its original source. These passages of 250 or more words were the ones used in this study. The letters A, B, C, and D were assigned arbitrarily to the passages within each difficulty level for identification purposes.

Cloze Tests: All five forms of a cloze test were made from each passage by selecting words 1, 6, 11, etc. to make the first form, words 2, 7, 12, etc. to make the second, and so on until all of the five forms possible had been made. Although deletions were made over the entire passage, only the first 50 items were scored. Responses were scored correct only when they exactly matched the words deleted, but spelling errors were counted correct if the response was otherwise correct. These scores were expressed as percentages. The reliabilities based on the split halves of these tests are shown in Table 5. These reliabilities are each based on the scores of 50 students drawn in equal numbers from grades 3 through 12.

Completion Tests: The completion test made from each passage consisted of 20 short-answer completion items. The items were made using roughly the same procedure used in the preceding study, but with two departures. In the preceding study the wh- phrase in the stems of a few of the items could not be replicated because the wording of the wh- phrase was left to some degree to the judgment of the item writer. In the present study an effort was made to reduce this problem by forcing a single wh- word into these slots whenever doing so did not make the question unanswerable. The result was, perhaps, to make the items more replicable; but this was achieved at the expense of obtaining items that were somewhat awkward grammatically. However, examination of the responses to these questions failed to reveal that they caused the students any special difficulty.

The second departure was to reembed constituents that were in questions that were not selected for use in the tests. Because longer passages were used in this study than in the last, more items were usually generated than were required for the tests. The constituents that had been disembedded to form these unused questions were reembedded in the appropriate stems.

The order of the items was randomized when the test was assembled to reduce prompting effects that might have arisen had the questions been placed in the same order as the relevant sections of the text. Each question was followed by two full lines of underlined blank space in which the student was told to write his response. The questions were scored using the same scoring procedures that were used in the preceding study. The scores were expressed as percentages. The reliabilities of each completion test when it was used as a pre-reading and a post-reading test are shown in Table 5.

TABLE 5

Reliabilities of Cloze, Pre-Reading, and Post-Reading Tests
for the Four Passages (A, B, C, and D) at Each Difficulty Level

Difficulty Level	Cloze Test				Pre-Reading Test				Post-Reading Test			
	Passage				Passage				Passage			
	A	B	C	D	A	B	C	D	A	B	C	D
1	.91	.75	.79	.87	.80	.56	.66	.76	.82	.78	.81	.78
2	.91	.77	.90	.78	.66	.46	.69	.84	.86	.76	.79	.54
3	.84	.62	.72	.76	.73	.49	.61	.12*	.81	.85	.86	.80
4	.79	.88	.92	.90	.60	.58	.70	.62	.80	.88	.82	.82
5	.93	.87	.80	.89	.63	.66	.46	.78	.82	.88	.79	.84
6	.90	.82	.90	.90	.72	.45	.57	.71	.86	.86	.91	.72
7	.89	.50	.69	.87	.45	.69	.30	.66	.75	.81	.52	.71
8	.80	.81	.90	.68	.06*	.22*	.63	.11*	.86	.81	.84	.21*

* Not significant at $p<.05$.

Preference Scales: Each student was required to rate a passage on ten scales. The first was the subject matter preference scale. The next three were the style, difficulty, and willingness-to-study scales, in that order. The student was asked to rate the passage on these three scales when it represented materials used as the regular textbook in one of his classes. The same three scales were then repeated two more times, once for obtaining the ratings when the passage represented materials used just for reference purposes, and again to obtain the ratings when the passage represented materials read voluntarily by the student. The four different scales are shown in Figure 7. The instructions for these scales were read orally to the students.

The instructions for the first scale, the subject matter preference scale, told the student *Look at question number 1. It asks how well you like to learn about the subject matter talked about in this book. On this question, do not try to rate the book, itself. Rather, we want to know just how much you like to learn about this subject, the kind of things this book talks about, regardless of whether you learn it by talking to others, watching T.V., reading a book, or listening to a teacher.* The instructions also explained the meanings of the numbers on the scale as represented by the verbal labels below each number. The student was instructed to indicate his response by circling the appropriate number. The instructions for this item and other items were repeated for any student who requested it.

The instructions for the style preference scale stated *How interestingly do you think this book was written--not what the author talked about but how he talked about it--if you had to use it as the regular textbook in one of your classes.* When this scale was used to rate a passage as representing a reference book, the last phrase was replaced with the standard phrase *if you just had to use it as a reference book for looking up special things.* When the passage was being rated for its use as a voluntarily read book, the standard phrase used was *if you just had to use it as a library book you would read just for pleasure in your spare time.*

The instructions for the difficulty preference scale were *Your response on this scale will show how hard or easy you think this book is for you* (standard phrase). The appropriate standard phrase was then inserted depending on the use for which the passage was being rated. The instructions for the willingness-to-study scale were *How well would you like, or how willing would you be, to read this book* (standard phrase). Again, the appropriate standard phrase was inserted depending on the use for which the book was being rated. The instructions contained a preface to the three scales used to rate the passage for each use that stated *The next set of three questions ask you how well you would like to study this book* (standard phrase). The instructions also contained the usual statements on when to go to the next item and when to turn the pages.

Subject Matter:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

Style:

1	2	3	4	5	6	7
Very dull	Dull	A little dull	Neither dull nor interesting	A little interesting	Interesting	Very interesting

Difficulty:

1	2	3	4	5	6	7
Much too hard	Too hard	A little too hard	About right	A little too easy	Too easy	Much too easy

Willingness To Study:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

FIGURE 7. Scales used to obtain preference ratings of the passages. The last three scales were used to obtain ratings of a passage when it was being considered for use as a textbook, as a reference book, and then as a book that the student would read voluntarily.

Testing Design: A test booklet contained a test over each of the four passages at one level of difficulty. The first page contained blanks for obtaining identification data on the student. The second page contained a cloze test made from one of the passages. The student was read the standard instructions given for cloze tests of this type (Klare, Sinaiko, and Stolurow, 1971). A completion test made from a second passage appeared next. The student was told that this was a test to determine how much students knew about passages that they had not read and he was urged to attempt every item. This test was followed by presenting the text of a third passage and the preference rating scales. The student was told that he was going to read this passage and then rate the book the passage was taken from. He was asked to examine the scales while each was briefly described to him. Following this he was instructed to read the passage, and each rating scale was then administered. The student was permitted to refer back to the passage at any time during the administration of the preference scales. Finally, the booklet contained the text of the fourth passage. The student was told to read this passage carefully and then to turn the page and attempt to answer the questions in the test that followed. Then he was told that this test contained questions of the same kind he had been asked to guess on earlier, and that he was neither allowed to examine the test before or during the reading of the passage nor allowed to refer back to the passage after he had started taking the test. The time limits employed for the testing were based on a field trial in which the maximum times required by samples of students at various grade levels were determined and used as the time limits here.

The test booklets were assembled using a Latin squares design. The passage orders ABCD, DABC, CDAB, and BCDA were defined. The first passage in each order was administered as a cloze test, the second as a pre-reading test, the third was rated on the preference scale, and the fourth was administered as a post-reading test. Five versions of each order were made, one for each form of the cloze test over the first passage in that order. This produced a block consisting of 20 booklets at one difficulty level. A total of ten such blocks was made at each difficulty level, one block being assigned to each of the grade levels 3 through 12. This process was then repeated at each level of passage difficulty. The eight blocks assigned to a grade level were then collected into a stack and the booklets in that stack were randomly ordered. When the tests were administered, the test administrator took the test booklets from the top of the appropriate grade level stack and passed them out to the students, each student being given whatever booklet happened to be on top at the time he was reached.

In summary, this design involved 10 grade levels of students crossed by 8 levels of passage difficulty, with four passages nested within each difficulty level and five cloze test forms nested within each passage. Thus, there were 1600 different booklets in the design and only one booklet of each type. The student was completely nested within all factors and with no replications. Thus, there were 160 students for each grade level and 200 students for each difficulty level.

Students Tested: The 1600 students tested were all enrolled in the same school system in a middle-class suburban community. Although an attempt was made to secure students at all grade levels who were members of a homogeneous community, the homogeneity of the sample obtained is difficult to certify. Housing developments tend to differ in the price ranges of the homes and different price levels tend to attract different populations of residents, each of which is to some degree homogeneous with respect to the ages of the children in the families and the occupational and educational levels of the parents. Moreover, the enrollment boundaries of high schools, junior high schools, and elementary schools are seldom geographically coterminous. As a result, it is difficult to avoid some correlation between student ability and grade level. The students tested were selected in consultation with the school officials who were responsible for conducting the demographic and achievement assessments for the school district. The object of the selection procedure was to obtain a sample of students that was homogeneous across grade levels with respect to their intelligence test scores and the educational levels of the parents. The students were sampled by intact classroom units for ease in administering the tests.

Test Administration: The tests were all administered within a span of three weeks. Each test administrator was given a manual containing a detailed set of instructions for administering the tests and was trained in its use. The instructions read to the students and the other testing procedures were the same in all cases. However, when the tests were administered to the very young students, special pains were taken to assure that they understood the instructions and they were given help with spelling when they asked the test administrators how to spell any word. Most of the test administrators were graduate students specially employed and trained for this task. The tests were generally administered in the students' classrooms. Although the teachers were asked to remain present during the testing, they were asked not to help except to refer the student to the test administrator when the student wanted to ask a question. Minimum time limits were specified for the test administrator. However, the administrators were instructed to provide additional time if a student requested it or if the test administrator could see that some students were still working productively. However, since the minimum time limits were quite generous, the test administrators rarely reported exercising these options.

Score Adjustments

Four adjustments were made to the scores. First, the scores on all the tests and scales were adjusted to remove the differences between the passage means within each difficulty level. Second, the raw gain scores were corrected to obtain residual gain scores. Third, most of the other scores were transformed into their true scores. Finally, the scores on the difficulty preference scales were pivoted about the midpoint of the scale.

Adjustment of Scores Within Levels: Four passages having similar cloze means were used in order to eliminate the carry-over effects that result from giving students, within a short time span, two or more tests made from the same passage. The assumption was that, if the passages were nearly alike in difficulty, the scores could be treated as if they had come from the same passage. However, the passages could not be perfectly matched even for cloze difficulty. Line 3 of Table 7 shows that there were significant differences among the passages within difficulty levels on both the cloze and completion scores, and Line 4 of Table 6 shows that similar differences also occurred on the preference scales. Consequently, some variability around the regression lines might arise from this source of variation, and, moreover, these variances between passages would appreciably reduce the sizes of the multiple correlations.

These corrections were relatively simple to perform for each of the preference rating scales. The ratings of all students taking the scale on a given level of difficulty were used to calculate a mean for each passage and a grand mean of all four of the passages at that difficulty level. Each passage mean was then subtracted from the grand mean and this difference was added to the rating of each student who rated that passage. These operations were then repeated at each difficulty level for that scale and then the whole set of operations was repeated for the remainder of the scales. This operation is valid only if the regressions among the pairs of passages at a given difficulty level are parallel. This assumption could not be tested directly since each student rated only a single passage. However, when the mean ratings were calculated for the five students at each grade level on each passage and scatter plots made of the regressions of these means, the regressions appeared to be fairly parallel. This fact may be verified by Lines 5 and 8 of Table 6, which show that the difficulty-by-grade interactions were extremely small and, in most cases, did not appear to be statistically significant in spite of the fact that the degrees of freedom for these statistical tests were extremely large.

However, the completion tests and, to a lesser extent, the cloze tests were affected by ceiling and floor effects that made these assumptions more tenuous in treating those scores. Some of the tests exhibited marked ceiling and floor effects, and scatter plots of the grade level means showed that the regressions were not parallel. However, most of these effects were removed by performing a probit transformation on the scores. This transformation may be performed by looking up a percentage score in a table showing the area under a normal probability curve and finding the corresponding deviation score. Zero and perfect scores were assigned the value $1/3I$, where I is the number of items in the test. For convenience in data handling, negative signs and decimal points were removed first by adding 5 to each deviation score and then multiplying it by 100. Thus, for example, the percentage scores 10, 30, and 50 became the probit scores 372, 448, and 500, respectively. These scores were retained in the probit metric through the regression analyses and transformed

TABLE 6

Analyses of the Variances of the Preference Rating Scales

Source of Variation	d.f.	Subject Matter		Style		Difficulty		Willingness-To-Study	
		M.S.	F	M.S.	F	M.S.	F	M.S.	F
1. Difficulty, D	7	29.99	2.92*	102.01	4.64**	704.55	35.71**	54.74	3.10*
2. Grade, G	9	36.36	12.28**	195.66	25.64**	50.62	11.38**	204.61	28.30**
3. Use, U	2			297.24	177.99**	2.49	3.83*	208.50	160.38**
4. Passage, P(D)	24	10.28	3.70**	21.97	3.21**	19.73	6.09**	17.68	2.55**
5. DG	63	3.88	1.40*	12.58	1.84**	3.56	1.10	11.13	1.61**
6. DU	14			5.14	3.08**	2.55	3.92**	2.19	1.68
7. GU	18			8.22	6.68	1.17	2.66	5.82	5.88
8. GP(D)	216	2.96	1.06	7.63	1.11	4.45	1.37	7.23	1.04
9. UP(D)	48			1.67	1.42*	.65	1.62**	1.30	1.34
10. GDU	126			1.18	.96	.48	1.09	1.00	1.01
11. GUP(D)	432			1.23	1.04	.44	1.10	.99	1.02
12. Student, S(GPD)	1280	2.78		6.85		3.24		6.93	
13. SU(GPD)	2560			1.18		.40		.97	

* $p<.05$ ** $p<.01$

back to percentages only for purposes of plotting curves. Lines 4 and 5 of Table 7 show that the transformations were fairly effective. The variances attributable to these interactions are trivial in size when compared to any of the main effects. However, it should be noted that the probit transformation was possibly less suited to the metric of the completion tests than to the metric of the cloze tests. The adjustments to the scores within each difficulty level were then performed using probit scores throughout the analyses.

Residual Gain Scores: As discussed in connection with Study I, a student's raw gain score provides a systematically biased estimate of his true gain score. These biases were removed in the present study by calculating a residual gain score in the manner described by Cronbach (1970). The first step consisted in calculating a beta. This began by performing analyses of the variances of the pre-reading scores separately at each difficulty level. This permitted estimations to be made of the mean squares among students within passages and grades, $S(PG)$, and of the variances between test halves within passages and students, $H(PS)$. Next, the covariance, cov, was calculated between the pre-reading, T_1, and post-reading, T_2, test scores within the difficulty level. The beta was calculated by

$$(35) \qquad \beta = cov/2S(PG) - H(PS) \quad .$$

The term $S(PG)$ was multiplied by 2 to correct for test length. Then this beta was used to calculate a residual gain score, G_r, for each student by

$$(36) \qquad G_r = T_2 - \beta T_1 \quad .$$

These calculations were then repeated using the scores obtained at each of the remaining grade levels. Hereafter, these residual gain scores will be referred to as the residual gain scores, information gain, or simply as the gain score.

True Score Adjustments: The scores on the cloze tests and the ratings on the preference scales were transformed into their true scores. As discussed in the context of Study I, this transformation was necessary in order to remove biases in the regression curves that are due to regression effects on the cloze scores, the independent variable. However, regression effects on the independent variables can also bias the shapes of the curves by causing each score to regress by different amounts toward the group mean. These corrections were performed separately at each level of difficulty, pooling the tests at that level and using the formula

$$(37) \qquad X_t = r_{1I}X + (1 - r_{1I})\bar{X}$$

where X_t is the student's true score, r_{1I} is the reliability of the test, X is the student's observed score, and $\bar{X}$ is the test mean. The reliability

TABLE 7

Analyses of the Variances of Cloze, Pre-Reading, and Post-Reading Tests after Probit Transformation

	Source of Variation	d.f.	Cloze		Pre-Reading		Post-Reading	
			M.S.	F	M.S.	F	M.S.	F
1.	Difficulty, D	9	1,277,909.70	93.15**	964,603.29	15.61**	2,447,843.84	42.69**
2.	Grade, G	7	226,259.48	64.85**	120,779.14	36.98**	418,298.44	65.50**
3.	Passage, P(D)	24	13,718.99	2.21**	61,789.57	3.32**[a]	57,344.10	2.50**[a]
4.	GD	63	3,871.97	1.11	5,239.22	1.60**	9,039.86	1.41*
5.	GP(D)	216	3,489.02	1.22*	3,265.91	1.29**[a]	6,386.32	1.48**[a]
6.	Halves, H(PDF)	160	2,806.61					
7.	Halves, H(PD)	32			17,398.55	14.84**	19,978.09	14.39**
8.	GH(PDF)	1440	550.57					
9.	GH(PD)	288			1,302.30	1.11	1,331.13	.96
10.	Forms, F(DP)	128	6,215.54	2.21**				
11.	GF(DP)[b]	1152	2,861.77	5.20**				
12.	Student, S(GDP)	1280			2,410.47	2.05**	4,384.86	3.16**
13.	SH(GDP)	1280			1,172.72		1,388.38	

* $p<.05$
** $p<.01$

[a] These were quasi *F* ratios calculated and tested in the manner described by Weiner (1962, pp. 199-202).

[b] This term also represents the variance attributable to students.

of the cloze tests was taken from the corrected correlations between the split halves of the tests. The error of measurement of the preference scales was estimated by performing an analysis of variance of each scale at each level of difficulty and then taking the interaction of use-by-students within grades, *US(G)*, the estimate of the error of measurement for that scale under all three use ratings. The ratings on the subject matter preference scale were not corrected since no statistic could be found that would provide a reasonably acceptable estimate of the error.

Pivot of the Difficulty Preference Scores: The format of the difficulty preference scales was set up in a manner that made scores on it awkward to use in their original form in the criterion selection model. While the ratings did represent a continuous scale of the ease of a passage, they did not represent a continuous scale of preference for the passage. Rather, extreme ratings of 1 and 7 each indicated that the passage was not desirable and the rating of 4 indicated that the passage was at the level the student most preferred. Consequently, it was necessary to pivot the ratings about this mid-point. Thus, ratings of 1 to 4 remained unchanged while the scores 5, 6, and 7 were transformed, respectively, to 3, 2, and 1.

Regression Analyses

The methods used to analyze these data were essentially the same as those used in Study II. Step-wise regressions were performed for each dependent variable. The independent variables were the first three powers of grade and difficulty, the first eight powers of cloze, and all possible two-way and three-way cross-products of these terms. The object of these initial analyses was to examine further the possibility that grade and difficulty each contributed uniquely to the regressions with the independent variables. Since most outcomes of this analysis were highly similar to those in Study II, only the contrasts will be presented in this discussion. Table 8 shows the correlations and standard errors obtained from the three sets of equations that resulted from entering just the cloze terms, from entering just the cloze and grade terms, and from entering the cloze, grade, and difficulty terms.

It can be seen that grade added somewhat to the regressions with the variables obtained from the completion test scores, and that passage difficulty provided a somewhat larger contribution to these regressions. In each regression where grade terms were permitted to enter, they did so, but with only small effects on the accuracy of the predictions. Passage difficulty also exhibited relatively minor effects compared to the effects observed in Study II. These contrasts with the results from Study II suggest that the large effects associated with grade and difficulty in the earlier study may have arisen, at least in part, as a result of the unreliability of the cloze tests in that study. That is, both the student's grade and the difficulty of the passage provide fairly good indexes of some of the same abilities as are measured by cloze tests, and

TABLE 8

Multiple Correlations and Standard Errors of the Regressions Involving Cloze, Grade, and Difficulty

Dependent Variable	Cloze		Cloze and Grade		Cloze, Grade and Difficulty	
	R	S.E.	R	S.E.	R	S.E.
Completion Test Scores						
Pre-Reading	.74	38.20[a]	.74	38.16[a]	.80	33.83[a]
Post-Reading	.77	58.71[a]	.78	58.06[a]	.84	49.49[a]
Information Gain	.68	75.05[a]	.69	74.48[a]	.77	65.46[a]
Preference Ratings						
Subject Matter	.11	1.92	.28	1.86	.30	1.84
Style						
Textbook Reading	.17	1.21	.33	1.16	.37	1.15
Reference Reading	.20	1.20	.29	1.17	.31	1.16
Voluntary Reading	.11	1.31	.39	1.22	.41	1.20
Difficulty						
Textbook Reading	.42	.75	.42	.75	.47	.73
Reference Reading	.34	.74	.37	.73	.42	.71
Voluntary Reading	.33	.76	.35	.76	.39	.74
Willingness-To-Study						
Textbook Reading	.11	1.30	.33	1.23	.34	1.23
Reference Reading	.18	1.26	.31	1.22	.32	1.22
Voluntary Reading	.10	1.35	.41	1.24	.42	1.24

[a]These standard errors are expressed in the probit metric.

in Study II they may have provided the more reliable measures of those abilities. Consequently, their large unique contributions may have arisen from this possibility. Thus, when the present study increased the reliability of the cloze tests relative to the reliabilities of the other measures, some of the contribution of passage difficulty and grade to the regression vanished. The terms entering these regressions did show that grade interacted fairly strongly with cloze in predicting the post-reading and information gain scores and it was, therefore, retained in the criterion selection model. However, because of its probable redundancy with the cloze score, the passage difficulty variable was still not incorporated into the model.

Grade played a much clearer role in the regressions of the preference ratings. It had a marked effect on increasing the correlations with all but the difficulty preference scales. Again, passage difficulty had only the effect of adding a small and fairly constant amount to the regressions after cloze and grade had been taken into account, indicating that grade level must be retained in the criterion selection model.

Plots of the Regressions

The regressions calculated for use in the model were also performed with the step-wise procedure. Just the terms involving cloze and grade were permitted to enter these equations. The criterion for entering these equations was that a term's partial correlation exhibit an F ratio (not the probability level of the F ratio) of at least .01, and the criterion for deleting a term from the equation was that this F ratio be below .005. These equations were then plotted against the column means for each grade level as a check on their fit. Table 9 shows the equations that resulted, their multiple correlations with the dependent variable, and their standard errors. When these standard errors were compared to those in Table 8, it was seen that the inclusion of more than just the statistically significant terms did not increase the standard errors of measurement. The curves shown in the graphs that follow were plotted only within the range observed to contain at least 49.5 percent of the scores on either side of the means for each grade level. Hereafter this will be referred to as the observed range.

Pre-Reading Test: Figure 8 shows a plot of the regression of the pre-reading scores on the cloze scores for each grade level of student. The chief effects observed in this regression were that there was a regular correlation with performances on the cloze tests and that grade level had a relatively small effect on the regressions. Grade level influenced students' scores only up to about grade 6.

There do not seem to have been any studies reporting an effort to explain student successes in guessing the answers to completion tests, at least not studies relevant to the effects of interest in these regressions. However, it can be seen that many students performed the task quite

TABLE 9

Equations Plotted and Also Used
in the Criterion Selection Model

(R = .74; S.E. = 38.13)

1. Pre-Reading Completion Scores = $-243.93G - 29.225C + .196C^4 - .0001242C^8 + 243.77G^2 + 544.4G^3 - 275.74CG^3 + 8.5881C^3G^3 + .7356C^4G - .05316C^6G^2 - .007936C^6G^3 - .00001136C^8G + .0007939C^8G^2 + 445.92$

(R = .78; S.E. = 57.98)

2. Post-Reading Completion Scores = $-588.24G - 16.247C^2 + .04242C^6 - .0008374C^8 + 1416G^2 - 429.49CG^2 + 16.393C^3G + .4107C^5G^2 - .1462C^6G + .0005369C^6G^3 - .0079C^7G^2 + .002441C^8G + 523.6$

(R = .69; S.E. = 74.53)

3. Information Gain Scores = $- 69.559G - 290.45C + 10.507C^3 - .007479C^7 + .0006534C^8 + 114.44G^2 + 54.76CG^3 - 25.188C^2G^3 + .6008C^4G^3 + .0001147C^8G - .0002042C^8G^3 + 838.87$

(R = .29; S.E. = 1.85)

4. Subject Matter Preference Ratings = $- 14.748G + .8766C + .001386C^5 - .00001306C^8 + 25.169G^2 - 10.361G^3 - 1.7347CG + .1915CG^3 + .0005736C^6G^3 - .0002258C^7G^2 + .00002921C^8G + 5.1297$

(R = .34; S.E. = 1.16)

5. Style Preference Ratings (Textbook Reading) = $- 9.5649G + .05538C^2 + .000059042C^7 - .00001232C^8 - .306G^3 + 2.8666CG^2 - .07572C^3G^3 - .001854C^5G + .0004407C^6G^3 + .00001449C^8G - .00001353C^8G^2 + 5.789$

(R = .30; S.E. = 1.17)

6. Style Preference Ratings (Reference Reading) = $- 6.0439G + .06572C^2 - .000001433C^8 + 1.4407CG^2 + .09189CG^3 - .03406C^3G^3 - .00006323C^6G^2 + .0000179C^7G^3 + 5.0903$

TABLE 9 (Continued)

(R = .39; S.E. = 1.22)

7. Style Preference Ratings (Voluntary Reading) = $-10.914G + .05029C + .0003794C^5 + .00003789C^6 + 9.3442G^2 - 1.1089G^3 - .8135CG^2 + .1827C^2G + .00004955C^6G^3 - .00008756C^7G + .00001603C^8G^2 - .0000072l3C^8G^3 + 5.7156$

(R = .43; S.E. = .75)

8. Difficulty Preference Ratings (Textbook Reading) = $.5922G + .03607C^3 - .0004054C^6 + .0000078llC^8 - 1.2921G^2 + .9912G^3 + .03172C^2G^2 - .00369C^4G^3 - .0002501C^5G + .0001781C^6G^2 - .000003423C^8G + .0000003501C^8G^3 + 2.0309$

(R = .38; S.E. = .73)

9. Difficulty Preference Ratings (Reference Reading) = $1.6323G + .4854C - .000151C^6 + .000004444C^8 - 2.3496G^2 + .8908G^3 + .0291C^2G + .04285C^2G^3 - .002905C^4G^3 - .00000646lC^8G + .000006898C^8G^2 - .000001218C^8G^3 + 1.2592$

(R = .36; S.E. = .75)

10. Difficulty Preference Ratings (Voluntary Reading) = $2.6339G + .7051C - .0004706C^6 + .00008228C^7 - 4.3988G^2 + 2.2578G^3 + .03862C^2G - .01715C^2G^3 - .000023l2C^6G^3 - .000009445C^8G + .00001036C^8G^2 - .000003168C^8G^3 + .3636$

(R = .34; S.E. = 1.24)

11. Willingness-To-Study Preference Ratings (Textbook Reading) = $-12.787G + .02021C^3 + .00005748C^6 - .000005455C^8 + 15.643G^2 - 6.7656G^3 + .08569C^2G^3 - .007263C^4G + .0001863C^5G^3 + .00008699C^7G - .00008705C^7G^2 + .000004382C^8G^3 + 5.8793$

(R = .32; S.E. = 1.22)

12. Willingness-To-Study Preference Ratings (Reference Reading) = $-5.079G + .03407C^3 - .000005513C^8 - 1.3847G^3 - 1.7203CG + 3.6122CG^2 - .1003C^3G^3 + .003492C^5G^3 - .0001856C^6G - .0003234C^6G^2 + .00001577C^8G - .000008297C^8G^2 + 5.2005$

TABLE 9 (Continued)

(R = .41; S.E. = 1.24)

13. Willingness-To-Study Preference Ratings (Voluntary Reading) =

$$-5.4321G + .7329C + .00003743C^{7} - .0000005559C^{8} + 1.4815G^{3} + .4348CG^{2} - .01887C^{3}G + .004379C^{4}G^{2} - .002501C^{4}G^{3} - .00018C^{6}G + .0000046830C^{8}G^{2} - .00000164 3C^{8}G^{3} + 2.8677$$

---

well; over 70 students, for example, answered 50 percent or more of the items correctly. This regression and the regression involving post-test scores provide some evidence that helps us to understand the students' successes at guessing. A preliminary analysis might list language redundancy, the information contained in the question stems in the test, and the student's prior knowledge of the contents of the passage as potential sources of information for this process. The completion item provides information about the part of speech and case membership of the response. For example, the wh- phrases what kind of, when, and for whom are regularly used to replace attributives, time adverbials, and benefactive human nouns, respectively. Students seem to take advantage of cues of this sort in selecting their responses. When error responses on cloze tests were examined (Bormuth, 1965) it was found that students rarely made responses that did not fit the syntax of the cloze blank. Inspections of the errors on the completion items in this study showed similar results; that is, the responses nearly always fit the syntax of the sentence if it were used to replace the wh- phrase and performed the same case function as the deleted phrase. The effect of this cue, then, would be to restrict the range of alternative responses possible to an item and thereby increase the probability of a correct response.

Similarly, the student's prior knowledge of the content discussed in a passage and the information about the passage that he is able to acquire from the stems of the questions in the test may serve to further select the correct response. The studies by Thomas J. Johnson (1971) and Paul E. Johnson (1969) have demonstrated that one of the effects of gaining knowledge about a topic is to increase the strengths of certain patterns of associations among the words used to discuss the subject. Thus, it could be speculated that a student might formulate his response to an item by associating a number of concepts with the words in the item stem, and then selecting from among those associates a response that fits the syntactic constraints signalled by the wh- phrase.

There is evidence for the claim that the guessing process has at least two components. This is seen in the fact that grade level had an

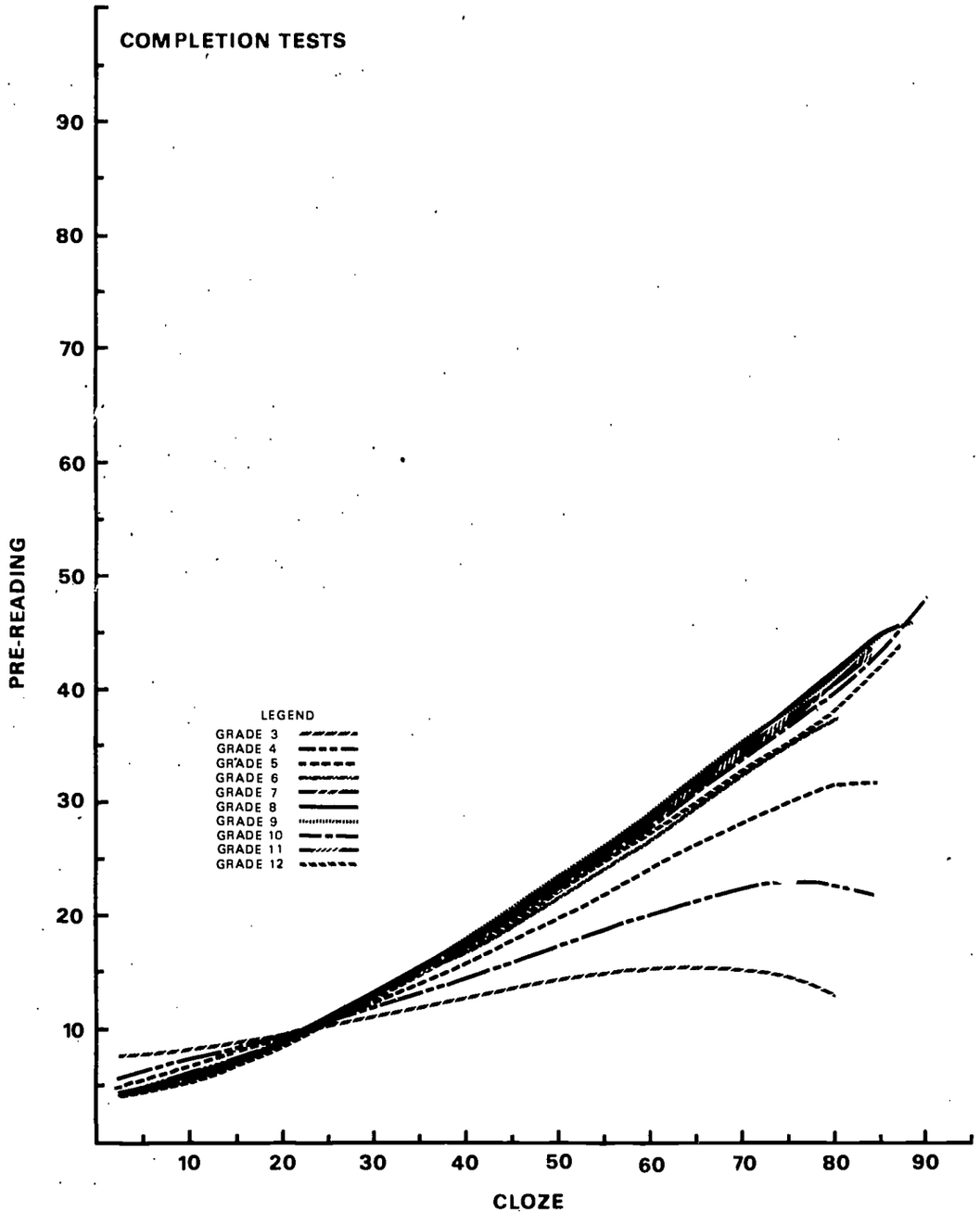


FIGURE 8. Regression of pre-reading completion test scores on cloze scores. It was speculated that the differences between the curves for the grade levels represented the development of a student's ability to take advantage of the semantic and syntactic information in the test itself for inferring the answers to the questions. Whatever this process is, it is apparently mastered by about grade 6.

effect only until the student arrived at grade 6. At that point, the students appeared to have mastered some component of the process. Moreover, the component that produced this grade level effect seems to be fairly independent of the processes involved in answering cloze items. If this process were dependent on cloze, there would have been no grade level effect observed in this regression.

It can only be speculated what the component producing the grade level effect might be. Perhaps it represents the gradual mastery of a strategy for obtaining information from the test itself, that is, a test wiseness such as an increase in the ability to utilize the syntactic information afforded by the wh- phrase of questions. It seems less likely to represent an increase in the ability to comprehend language in general and particularly that language in the test items. The evidence for this assertion will be taken up in discussing the regression of the post-reading scores. It also seems unlikely that this grade level effect represents increments in general knowledge about the contents of passages; this would be expected to show growth beyond grade six and also to affect cloze and completion test scores in about the same ways. Nor does the grade level effect seem likely to represent increments in ability to cope with the spelling and handwriting tasks involved in making the responses; effects of this sort would also be likely to influence the cloze and completion tests about equally and, therefore, produce no grade level effect. In any case, the nature of this grade level effect was such that it, alone, would force the information gain curves to differ in shape for different grade levels, making it necessary to select different performance criteria for students at some grade levels.

Post-Reading Test: Figure 9 shows a plot of the regression of the post-reading scores on the cloze scores for each grade level of student. The chief effects observed in this regression were the absence of a correlation between the heights of the intercepts and the grade levels of the students, the appearance of ceiling and floor effects on the completion tests, and the grade-by-cloze interaction that resulted in the curves fanning out to some extent. In Study II the heights of the intercepts had correlated with the grade levels of the students and a follow-up study provided some evidence that this was due to a carry-over effect from administering the same test as a pre-reading and post-reading test. The fact that the effect failed to occur in this study confirmed the results of the follow-up study. The curves also show the anticipated floor and ceiling effects on the completion tests.

The fanning out of the curves on these tests appeared to be an exaggeration of the grade effects observed in the regression involving the pre-reading scores. That is, the effect was greatest at the lower grade levels and gradually diminished at the upper grade levels. Consequently, the same speculations may also be appropriate for interpreting the fanning out observed in this regression. If this interaction were no more than an exaggeration of the one observed in Figure 8, then it could be

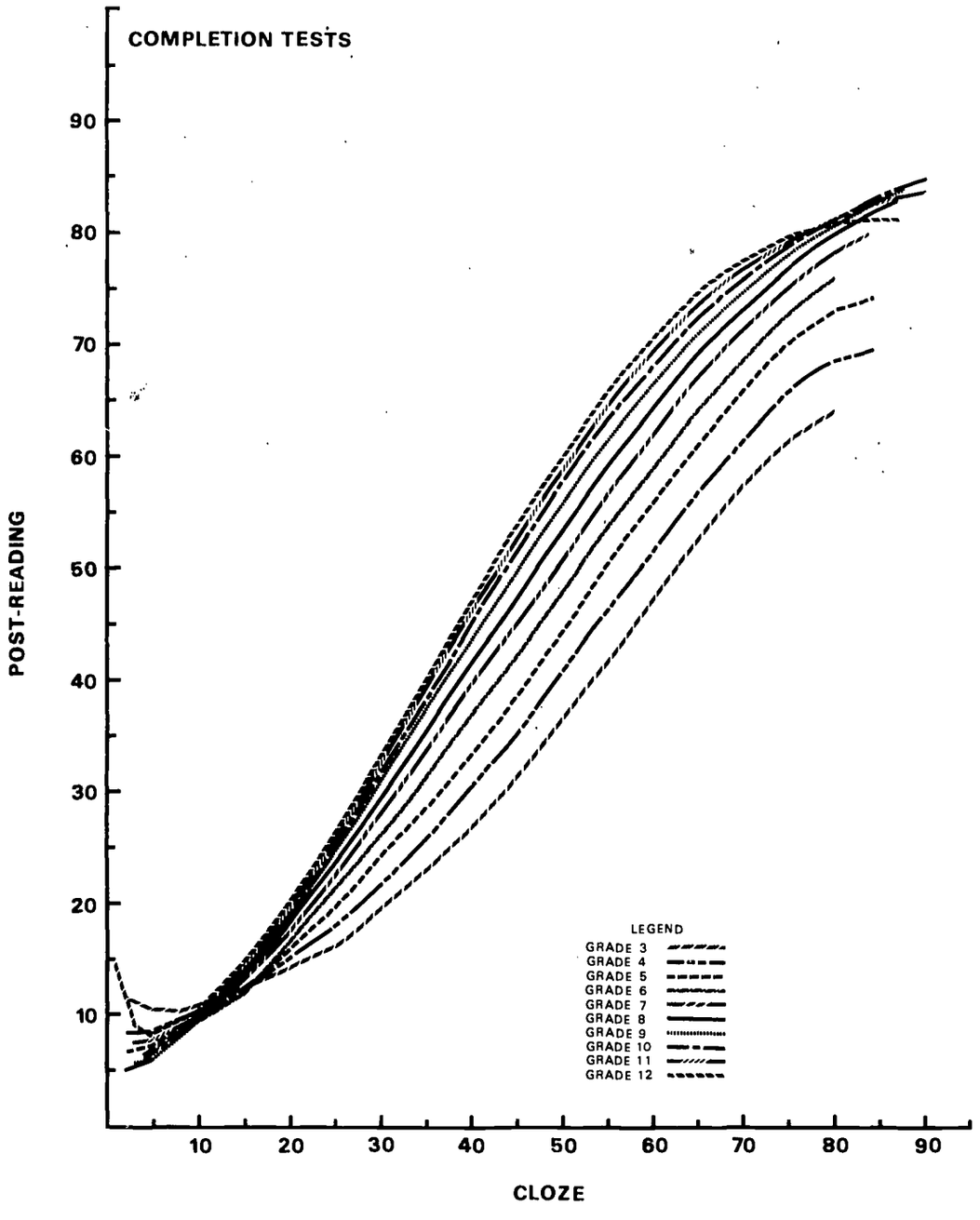


FIGURE 9. Regression of post-reading completion test scores on cloze scores. It was speculated that the differences between the curves for each grade level were caused by a combination of effects. The students were gradually developing both the ability to utilize cues in the tests themselves for inferring the correct responses and the ability to anticipate and focus on the type of passage content likely to be tested.

interpreted as supporting the argument that it was due to the development of test-taking strategy. If the effect were due merely to an increase in general language comprehension, this would provide the older students with some advantage in understanding the items on both the pre-reading and post-reading tests, but with greater advantage on the post-reading test, where they had the entire passage to study as well as the test items. However, if this explanation were accepted, then a new theory would have to be sought to explain why the same skills did not have a similar effect on the cloze tests, where 80 percent of the words of the passage are preserved as well as the ordinal relationships among the sentences and paragraphs of the passage. Ordinarily an increase in comprehension ability would be expected to influence scores on the two kinds of tests in the same way and thereby reduce or entirely eliminate the grade effect and the grade-by-cloze interaction.

This leaves the alternative explanation, the development of a test-taking strategy, as the more likely explanation. Presumably, such strategies involve not only the skills appropriate for guessing the answers to the test items but also the attention-focusing skills of the type demonstrated by Rothkopf (1966). That is, the students seem to learn to expect certain kinds of content to be tested and to focus their attention on learning that content. Thus, grade level effects observed on pre-reading tests would be expected to increase when the student also had the passage, itself, to study. These test-taking strategies would not necessarily be appropriate for taking cloze tests, where the material tested is present as a part of the test itself. Consequently, a grade level effect might be expected to occur, and to occur primarily on the completion tests. If this interpretation is accepted, then Figure 8 could be interpreted as showing that the students reach maximum development of the skills used just to guess the answers to test questions by grade 6 and continue to develop in ability to focus attention on material likely to be tested until about grade 9 or 10.

Information Gain: Figure 10 shows the plot of the information gain scores on the cloze scores. The gain score axis has been exaggerated relative to the other graphs in order to show more of the details of these curves. This regression conforms fairly closely to the shape of the regression observed in Study I. That is, it exhibited regions of near zero slope in the lower and upper ranges of cloze scores. Moreover, it exhibited the anticipated decline in scores at the upper extreme of cloze scores, an effect that was not clearly evident in Study I. This decline in scores is interpreted merely as a mathematical consequence of subtracting pre-reading scores, which continue to rise in this region, from post-reading scores, which level off because of ceiling effects. As anticipated, these curves differ in shape and slope for each grade level. Consequently, it was necessary to incorporate the student's grade level into the criterion selection model when this variable was considered.

Subject Matter Preference Ratings: Figure 11 shows the regression of the subject matter preference ratings on the cloze scores. These

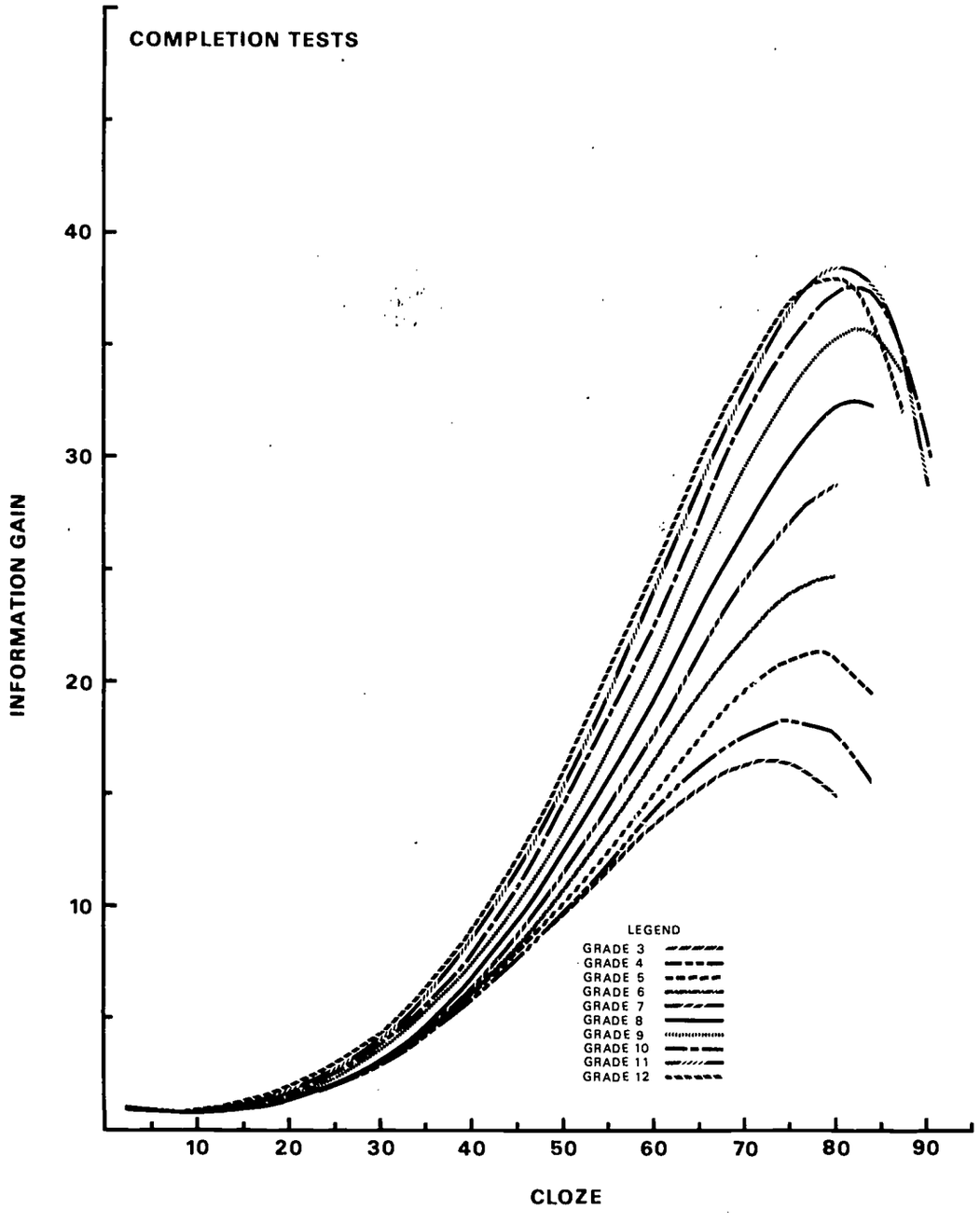


FIGURE 10. Regression of information gain scores on cloze scores. Since these curves differed in slope and shape for each grade level, it is necessary to incorporate the student's grade level as a parameter of the criterion selection model when considering this variable.

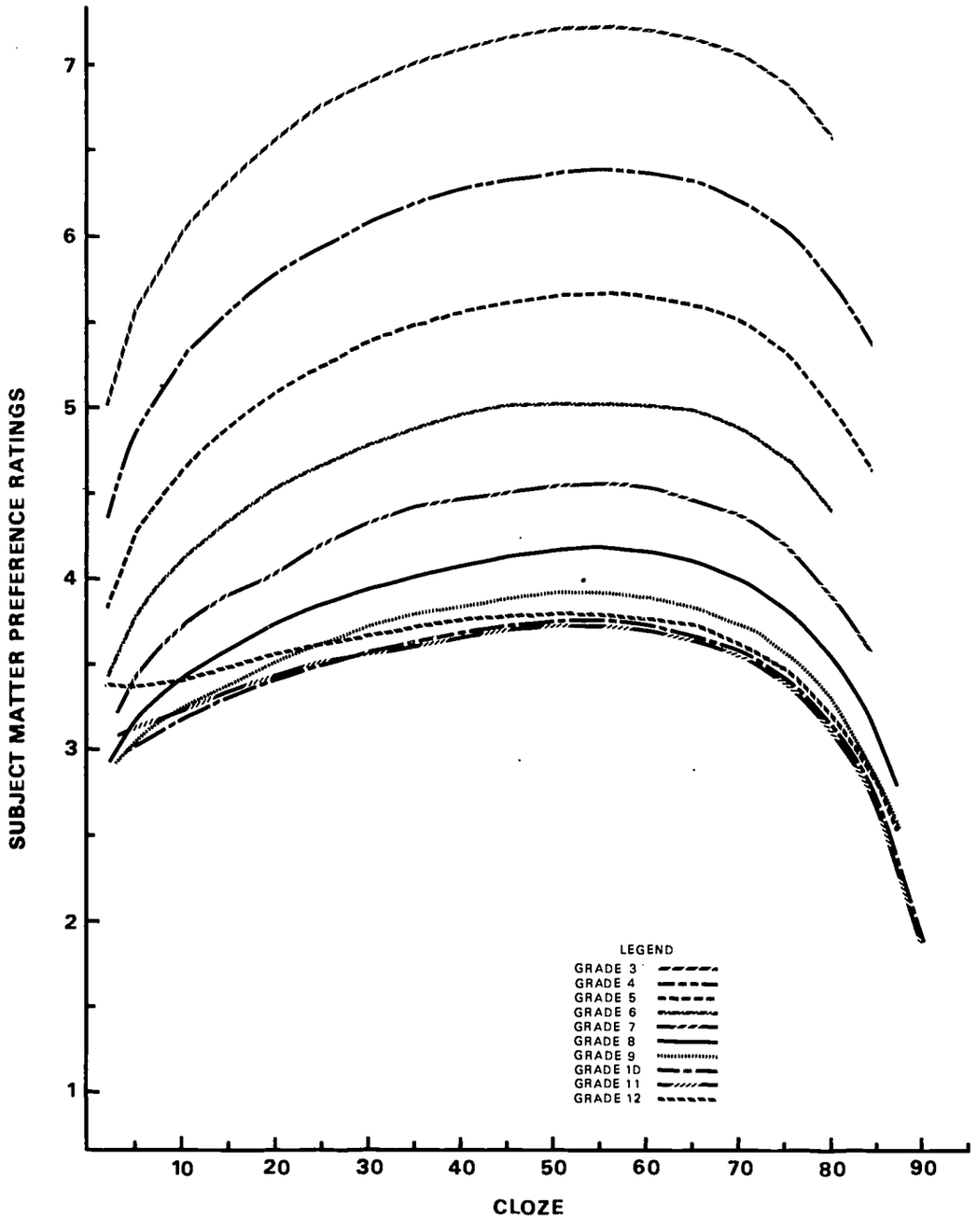


FIGURE 11. Regression of subject matter preference ratings on cloze scores.

curves show three major characteristics--a general decline in ratings as students reached higher grade levels, a tendency for students to give lowest ratings to passages at the extremes of cloze scores, and a tendency for the relatively low preference ratings of passages at low cloze scores to vanish as students reached higher grade levels. But perhaps the most compelling question is why these ratings should correlate with cloze scores at all when the rating scale supposedly measured preference for the topic regardless of how that topic was presented. As was pointed out in the discussion of Study II, the design of these studies deliberately confounds structural complexity with topic familiarity in a manner that hopefully reflects their confounding in naturally occurring instructional materials. While this may lead to more generalizable results, it can also lead to correlations such as this one. Since the passage difficulty dimension represented a scale of topical familiarity that also happened to correlate with cloze scores, the results can be explained as due, at least partially, to topical familiarity--and the regression on cloze scores as merely a by-product of the correlation between the topical familiarity and cloze difficulty of the passages. It is also possible that the students were not entirely able to abstract the passage content from other attributes of the passages and react just to passage content.

Thus, students in higher grade levels generally rated all passages lower because those students had become more familiar with the topics of all the passages. All students tended to rate relatively low the passages on which they obtained high cloze scores because the topics of those passages were highly familiar to those students. The passages on which the students achieved very low cloze scores, on the other hand, did not become more desirable to the students as the students reached higher grade levels. Rather, the remainder of the passages merely declined more in desirability relative to those passages as the students became increasingly familiar with the topics discussed.

This account, however, seems to leave unexplained the fact that younger students rated relatively low the passages on which they attained relatively low cloze scores. This can be explained in either of two ways. First, it could be conceded that the preference ratings might have been contaminated by the students' reactions to the structural complexity of the passages, thus permitting the effects of the Dember-Earl model to explain the decrease as being a result of the students' lower preference for materials of high structural complexity. There does not seem to be any evidence in these studies for excluding this explanation.

Second, this effect could also be explained by arguing that the relationship between topic familiarity and preference ratings follows an inverted V shape much like the relationship between preference ratings and structural complexity. This is not an unreasonable proposition since it is common knowledge that, while novelty is generally sought by people, extremely unfamiliar and novel forms tend to elicit avoidance and fear reactions. Such things are labeled as being "strange" and "weird," labels that not only denote novelty but also connote danger. If this explanation

were accepted, and again there seems to be no evidence here for rejecting it, then the decrease in the preference ratings by very young students on passages that were difficult for them indicated that the topics of those materials were so novel that they constituted some threat to the students. Although the author slightly prefers the latter interpretation of these results, it must be clear that the interpretations of all of the preference ratings are subject to several alternative explanations that appear to be equally acceptable.

Style Preference Ratings: Figures 12, 13, and 14 show the style preference ratings for textbook, reference, and voluntary reading, respectively, regressed on cloze scores. These curves were similar to each other and showed a marked similarity in shape to those obtained from subject matter preference ratings: There was a large effect due to grade level; ratings dropped sharply at the extreme ranges of cloze scores; and the curves varied systematically in shape across grade levels. These similarities would seem, at least superficially, to suggest that the students were rating essentially the same features of the passages regardless of the scale they were responding on. But some evidence tends to refute this proposition and to support the explanation that stylistic and subject matter variations had a fairly high degree of correlation in the passages rated and that, taken together, they did much to determine passage difficulty. Thus, it could be argued that, although the curves on different scales appear to be superficially similar, they may represent responses to quite different aspects of the passages.

Perhaps the strongest evidence for this interpretation may be seen in Table 6, where the analysis of variance showed that there was a large amount of variance in the style ratings that was attributable to the use factor. Here the subject matter of the passages was being held constant, so that the variance could not be attributed to subject matter variations in the passages. Yet, when use was varied, large variations were observed in the style ratings. The variations among the figures show the approximate forms of these contrasts. The subject matter preference ratings by upper grade students did not exhibit a sharp drop in the lower range of cloze scores, while the style ratings uniformly did so. Similarly, the main trend of the curves for the subject matter ratings is a positive slope at all grade levels, while the trends of those curves tend to be negative for the style ratings by the upper grade students.

These results also show that the criterion selection model should take use into account in identifying performance criteria. The clearest evidence is the sharp contrast between the shapes of the curves in Figure 14 and the curves in the other two figures. The curves for grades 3 and 4, for example, fail to show any decline in the upper range of cloze scores. At the same time, the curves for upper grade students tend to show relatively little variability in Figure 14 when they are compared to the curves for the same grade levels in the other two graphs. Consequently, the student's style preferences appear to depend to some

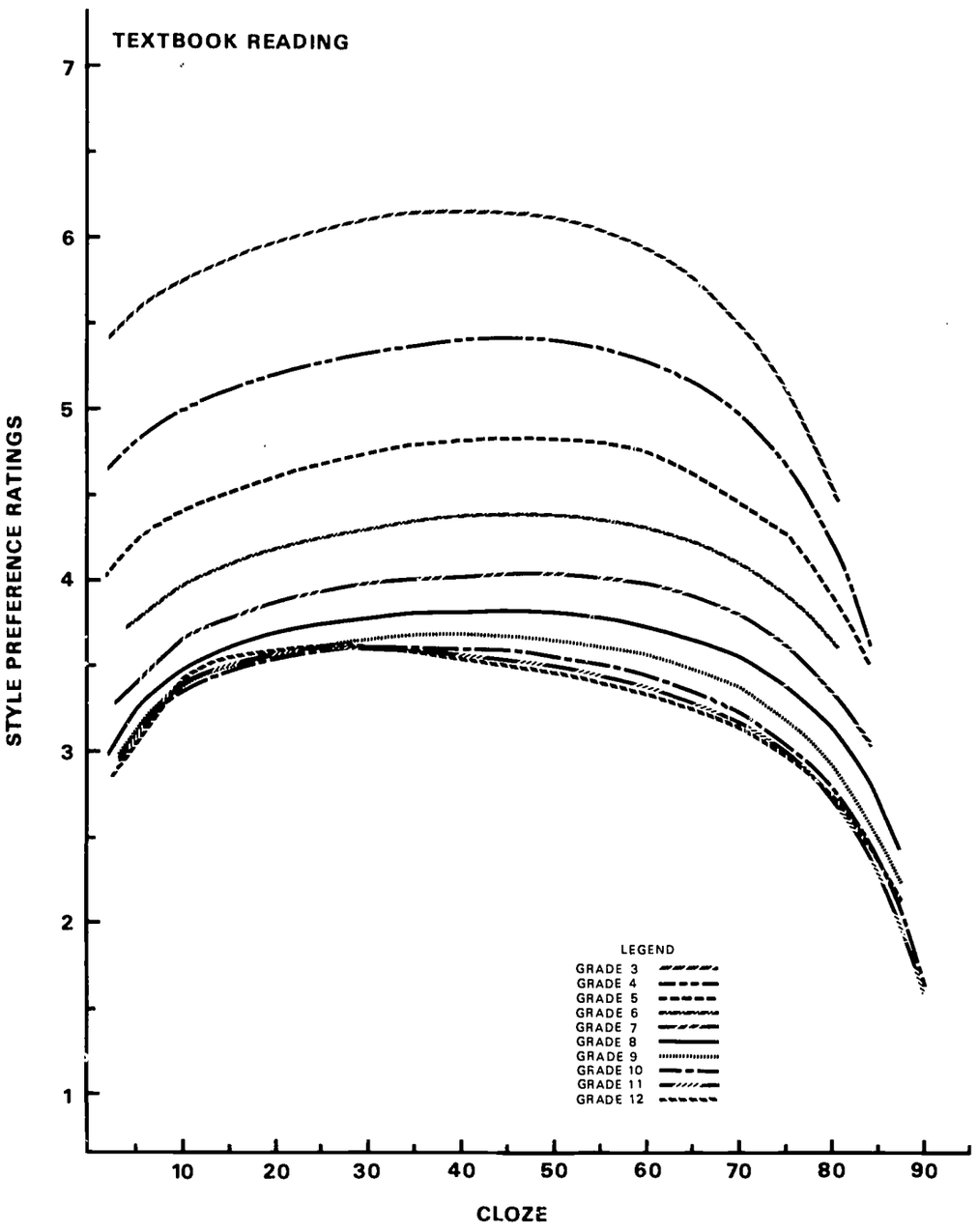


FIGURE 12. Regression of style preference ratings for textbook reading on cloze scores.

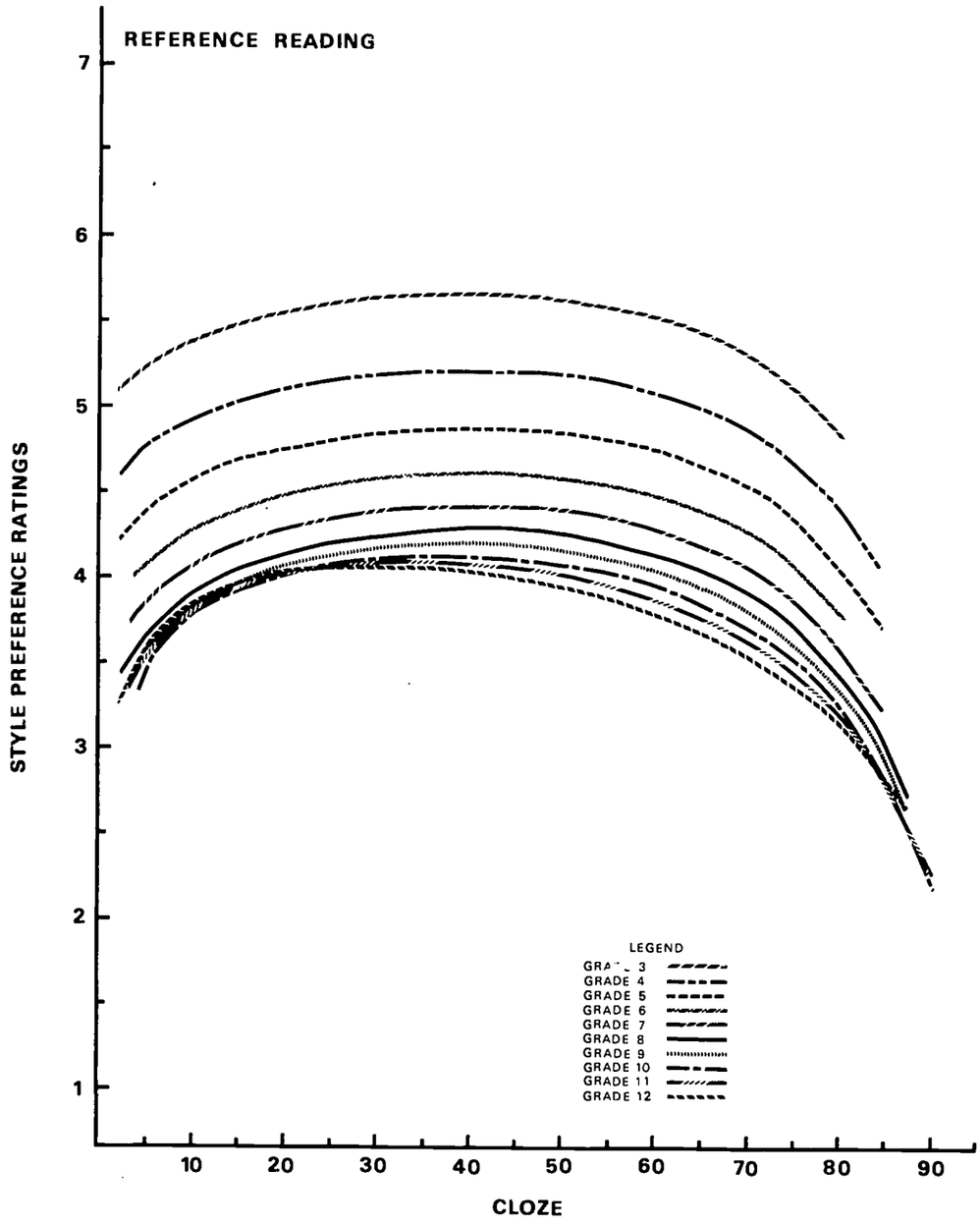


FIGURE 13. Regression of style preference ratings for reference reading on cloze scores.

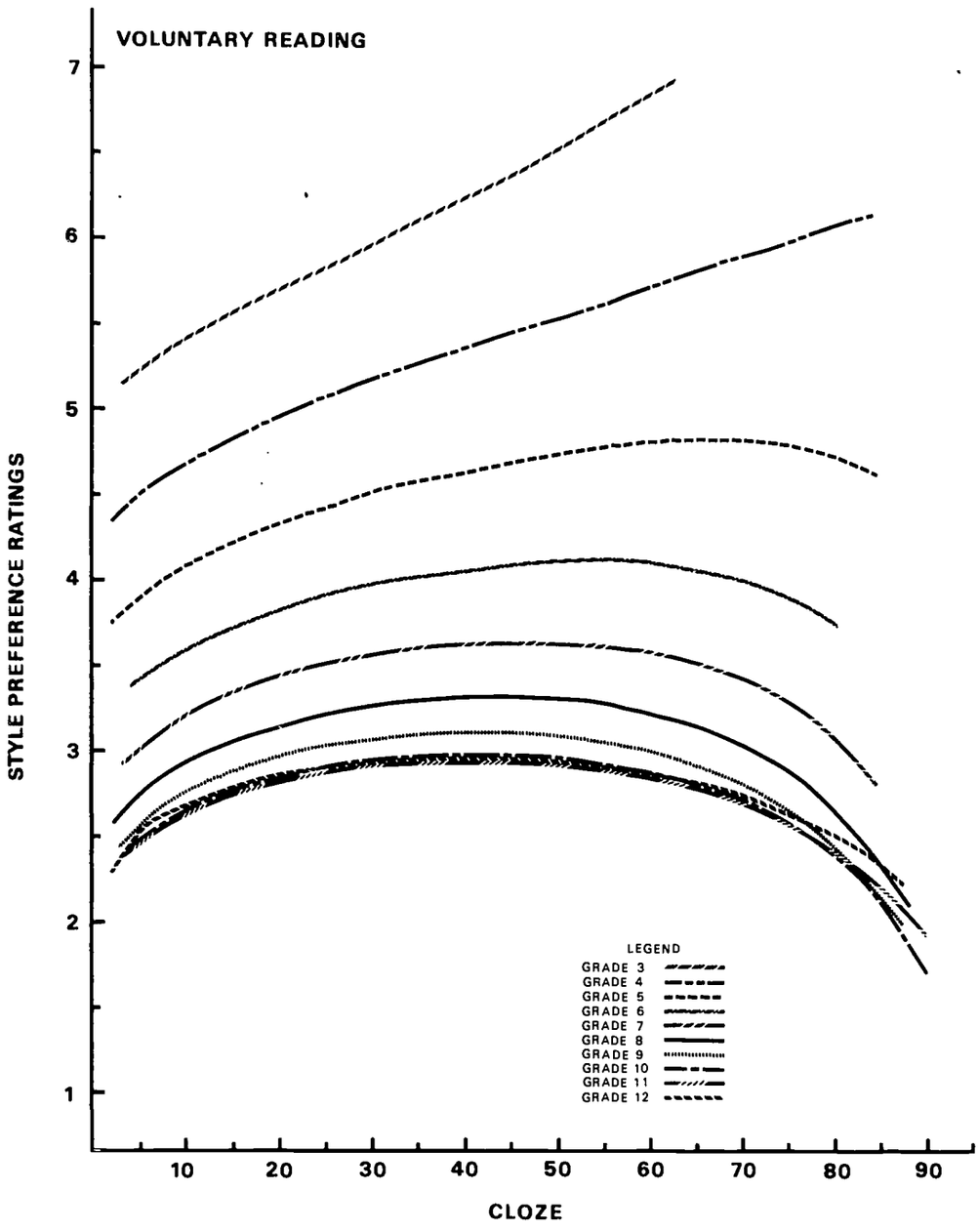


FIGURE 14. Regression between style preference ratings for voluntary reading on cloze scores.

extent on the use for which the material is being considered, and this, in turn, affects the preference rating assigned at a given level of cloze performance for that passage.

Difficulty Preference Ratings: Figures 15, 16, and 17 show the regressions of the difficulty preference ratings on the cloze scores for the three use conditions--textbook, reference, and voluntary reading, respectively. The results on these scales conformed fairly closely to expectations that would be consistent with the concept of the scale as a measure of the suitability of the content of a passage for the student. The ratings reached a maximum at a cloze score of about 10 to 20 percent on the cloze scale and then declined steadily as higher cloze scores were attained. This result fit neatly with the results on the completion tests shown in Figures 8, 9, and 10. At cloze scores below 15 percent the student was able to gain little or no information from the passage and so he would have no basis for rating the passages as being still more suitable for him. Consequently, these ratings would be expected to level off at this point. At higher cloze scores one would expect to observe a steady decrease in the ratings since the students obtaining cloze scores in that range were also able to demonstrate increasing amounts of prior knowledge of the content of the passages.

However, two effects were not anticipated. First, it was not expected that the students' ratings would drop sharply on passages that were below 15 percent on the cloze scale for them. It seemed reasonable to assume that the ratings would merely level off in this region. Since the passages at this level were almost uniformly incomprehensible to the student, it seemed clear that he would be unable to make systematic discriminations among them and would simply rate them much as he would the passages that he could understand slightly. Second, it was not anticipated that the curves of the younger students would exhibit the very steep increases in the upper range of cloze scores.

Interviews with some students given these rating tasks over several passages provided some grounds for speculating that both effects may have represented a switch in set on the task caused by a fear of inability to understand the materials. For example, when confronted with very difficult passages and asked to justify their ratings, many of the remarks were of the form "It's so hard I couldn't even understand what it was talking about." Thus, the rule being followed might be stated as: When the passage is so difficult that it cannot be understood, students ignore their knowledge of the topic of the passage and rate its understandability. Similarly, when the younger students were asked to rate the passages on the reference and voluntary reading scales, they often mentioned the fact that they would not have a teacher's help in these types of reading situations and therefore probably couldn't understand the materials. This might account for their tendency to rate the very easiest materials highest and all other materials lower in these use conditions.

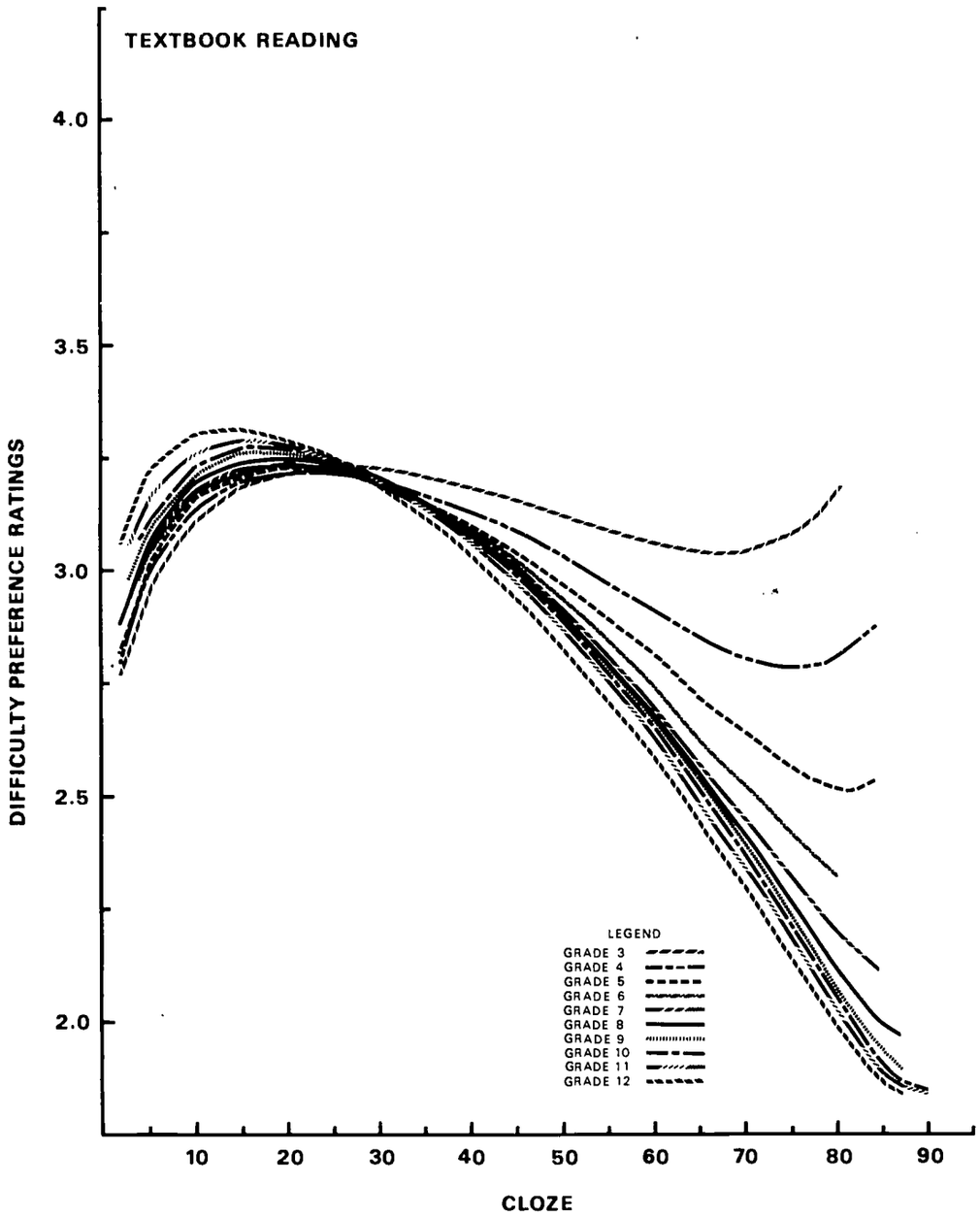


FIGURE 15. Regression of difficulty preference ratings for textbook reading on cloze scores.

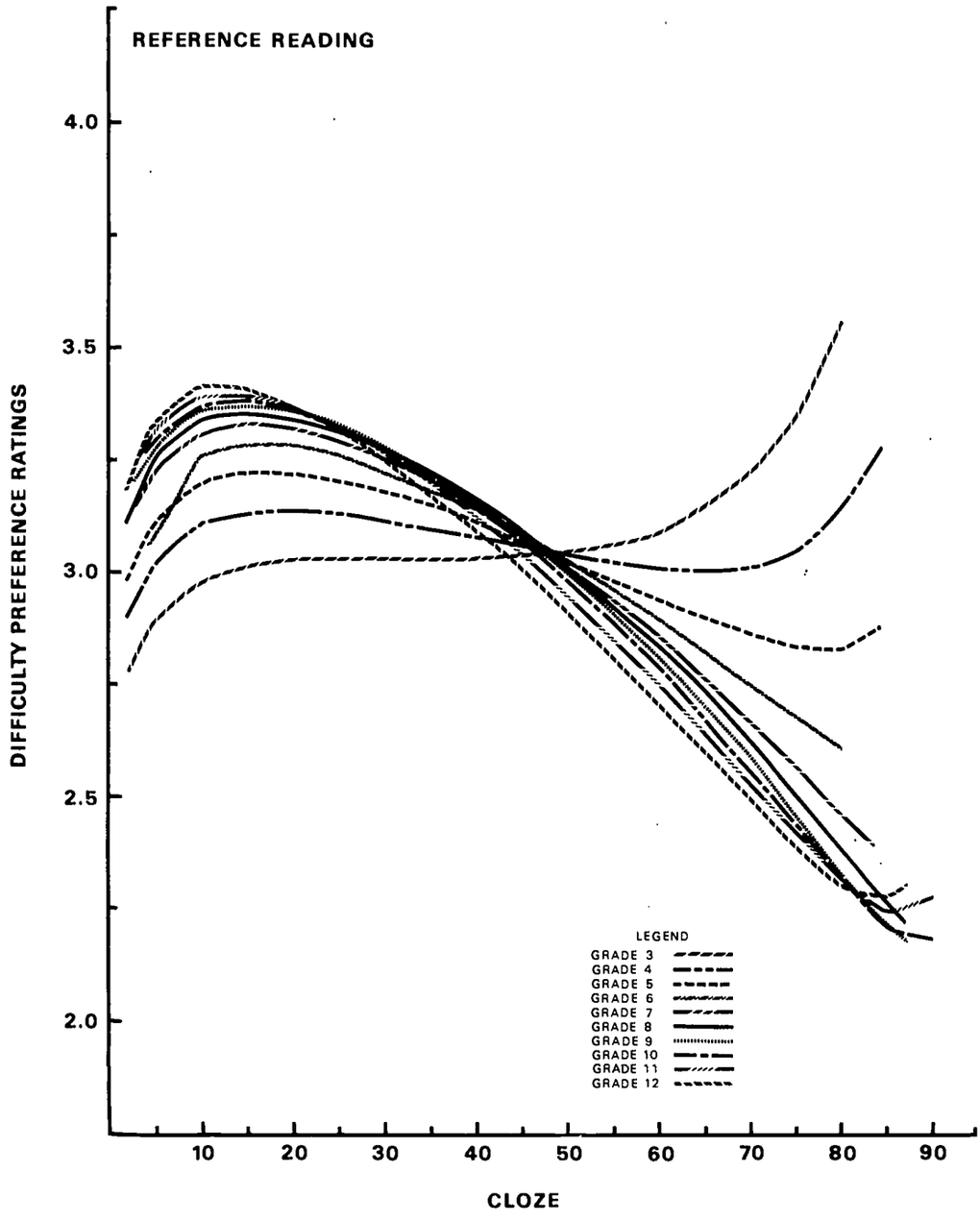


FIGURE 16. Regression of difficulty preference ratings for reference reading on cloze scores.

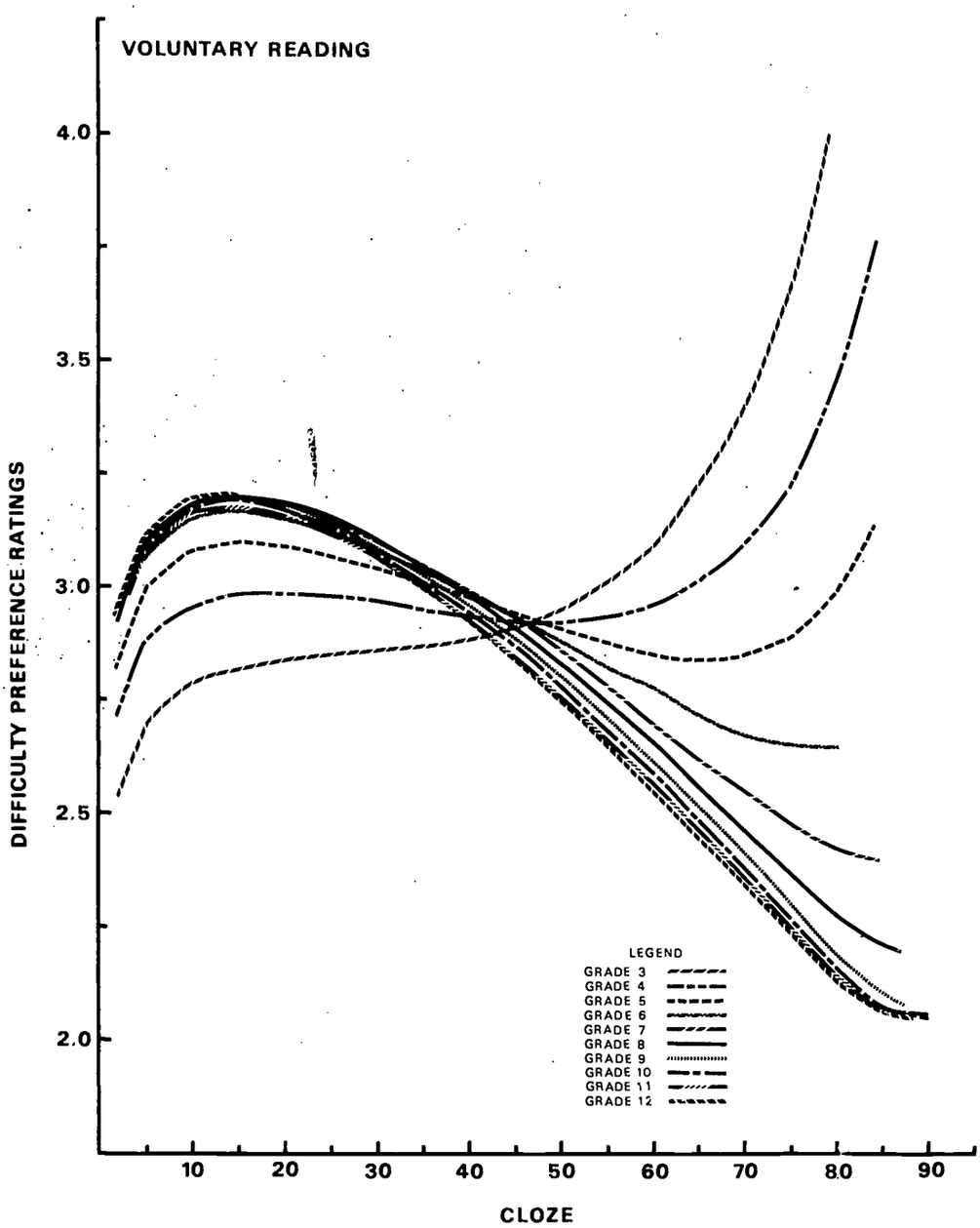


FIGURE 17. Regression of difficulty preference ratings for voluntary reading on cloze scores.

These regressions seemed to require that the criterion selection model take into account both the grade level of the student and the conditions under which materials are to be used when treating difficulty preferences. The grounds for this assertion are that the curves differed in shape for some grade levels and for each of the three use conditions.

Willingness-To-Study Preference Ratings: Figures 18, 19, and 20 show the plots of the regressions of the willingness-to-study preference ratings on cloze scores. The regressions show several common and interesting features. Each exhibited the fairly large effects due to grade level that have been interpreted as being due to the increasing familiarity of the older students with the content of the materials. There was also a tendency for the curves of students in the lower grades to peak at somewhat higher cloze scores than the curves for the students in the higher grades. When the same effect was observed in Study II, it was interpreted as being due to the facts that topical familiarity and structural complexity were confounded in the passages tested, and that the younger students found the topics of all passages generally less familiar than the older students had and therefore rated even the easier passages relatively high.

There were also important contrasts among the regressions. The curves for the textbook and voluntary reading conditions peaked at somewhat higher cloze scores than the curves in the reference reading condition, suggesting that the student perceived the reference and voluntary reading conditions as providing him with less teacher assistance in acquiring the information content of the materials and as, therefore, making more demands on his personal reading abilities. The fact that the third grade students' curve failed to decline at all in the voluntary reading condition, then, apparently indicated that they felt that they had little of the knowledge contained even in the easiest passages, and that the passage should present them with a minimum of structural complexity since they would have to rely mainly on their own reading skills in this use condition. However, as the students reached higher grade levels they may have felt that the structural complexity of the easier passages was below their preferred level and that the topics of those passages were already fairly familiar to them. Consequently, their curves exhibited a decrease in upper range of cloze scores.

Interpretation of the Preference Regressions as a Whole: It is tempting to think of the willingness-to-study variable as being the ultimate measure of student preference and to think of the other preference measures as constituting merely its components. This would lead to efforts to interpret the shapes of the willingness-to-study curves as being composites of the other curves. For example, the shapes of these willingness-to-study curves could be fairly accurately reproduced by taking a weighted sum of the information gain, the subject matter, the style, and the difficulty regression curves. In fact the willingness-to-study variance can be remarkably well predicted by a multiple regression

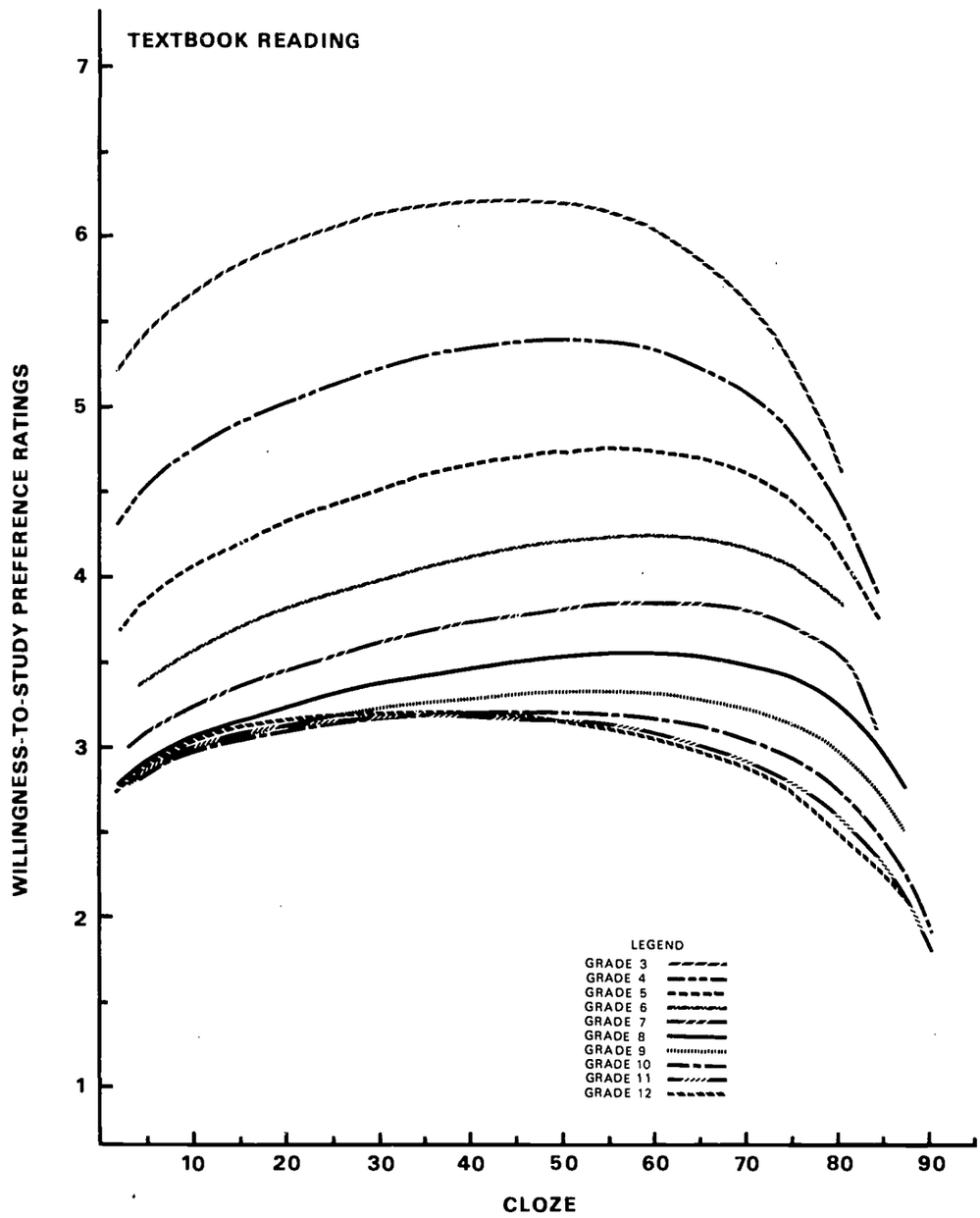


FIGURE 18. Regression of willingness-to-study preference ratings for textbook reading on cloze scores.

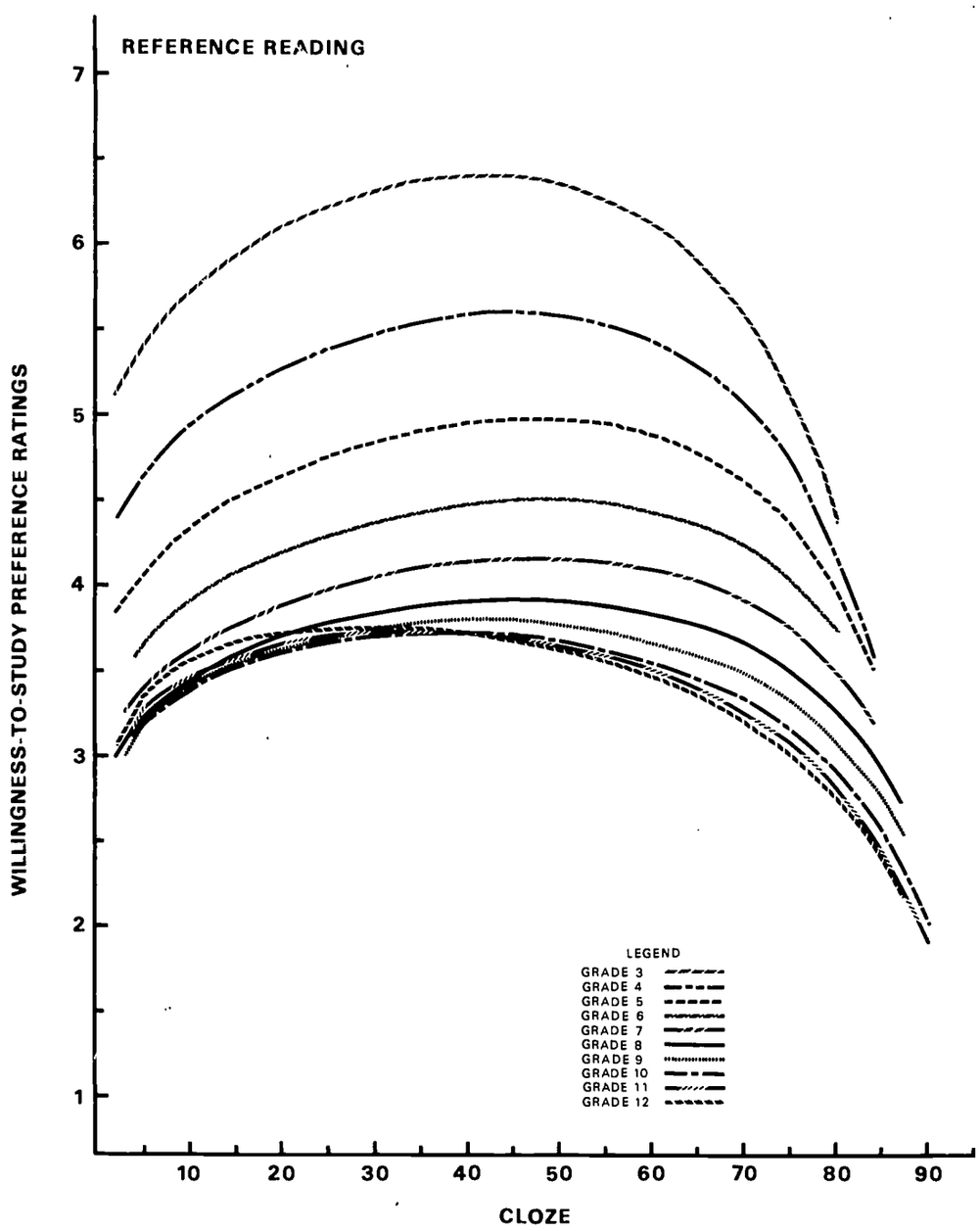


FIGURE 19. Regression of willingness-to-study preference ratings for reference reading on cloze scores.

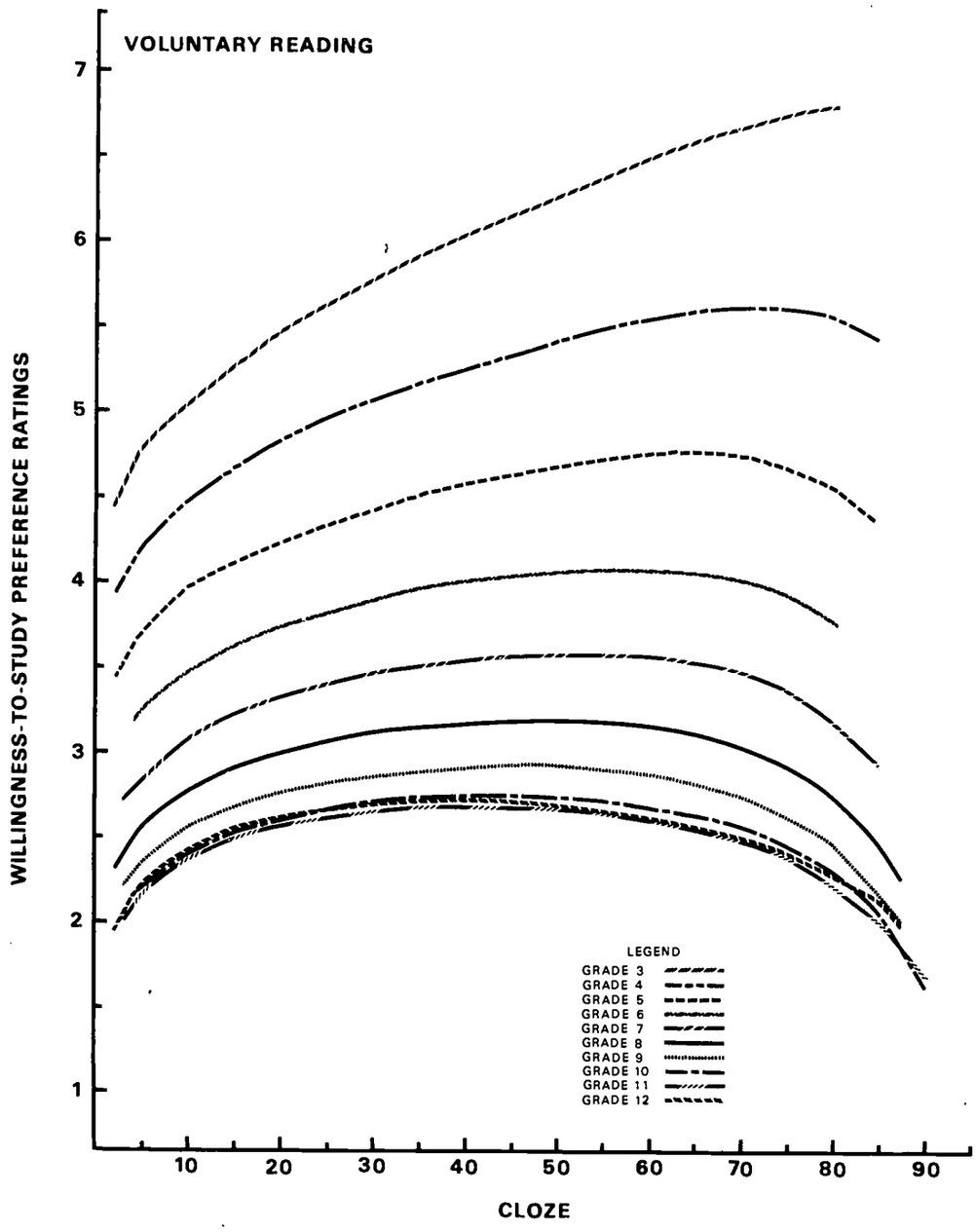


FIGURE 20. Regression of willingness-to-study preference ratings for voluntary reading on cloze scores.

involving these variables. However, such interpretations may be premature since they depend on a theory that relates these variables and at the present time no such theory exists. As a result, the curves are interpreted here primarily in terms of the Dember-Earl theory and the proposed extension to that theory. And, for present purposes, it seems more justified to treat each of these scales in the criterion selection model as constituting a value that should be treated independently in the model. As the theory of preferences improves, corresponding improvements can be made in the criterion selection model. In every case, however, the student's grade level and the use for which the materials were being evaluated altered the shapes of the regressions. Consequently, it seemed essential to take both factors into account when dealing with the preference ratings in the criterion selection model.

Summary and Evaluation

The primary object of this study was to obtain regressions that would be suitable for demonstrating the criterion selection model in its current state of development and for making at least a preliminary identification of a set of passage performance criterion scores. The other major objective was to do further exploratory investigations of passage preferences. The results could probably be accurately assessed as a limited success on all counts.

The shapes of the information gain curves appeared to agree reasonably well with the theoretical expectations and, since the multiple correlations were fairly high, it would appear that the test-making procedures used resulted in fairly replicable regressions from test to test. However, the evidence for their replicability is merely empirical, and considerable improvements will have to be made in those test-making procedures before it will be possible to make a persuasive logical case that the regressions are replicable. Moreover, these tests measure only what is commonly termed literal comprehension. Thus, the classes of items included in the tests will also have to be extended before the criterion selection model can be said to take account of all the important cognitive benefits the student can accrue from reading instructional materials.

The regressions obtained from the preference scales produced several useful results. It appeared possible to analyze preferences into subject matter, difficulty, style, and willingness-to-study components and still obtain interpretable results. Moreover, it appeared that the use to which materials are to be put makes a considerable difference in how the student rates them on each of these scales. This result added another dimension, the use dimension, to the criterion selection model. However, there seem to be relationships among these preference ratings that, to some degree, make them redundant. For example, it seems possible that the student's willingness to study may be completely determined by his preferences for the subject matter, [illegible],

and difficulty of the passages. If so, including it along with the others in the criterion selection model would have the implicit effect of giving it a double weighting in identifying a performance criterion. Consequently, these preference variables must be analyzed further before their proper treatment in the model can be determined.

STUDY IV

Purpose

A preliminary study was conducted to explore the relationship between rate of reading and cloze scores. These data seemed to be of sufficient quality that this regression could be included in this demonstration of the criterion selection model. It should be clear, however, that this study furnishes only a preliminary determination of the nature of this regression and that more elaborate studies must follow.

Procedures

The materials and procedures followed in this study were nearly identical to those in Study III. The same test materials and nearly the same instructions were used. The chief contrast in the testing procedures was that rate of reading was measured for the passage on which the students took the post-reading test. The student was given the usual instructions for reading the passage and taking the completion test that followed it, and then he was asked to write on the booklet the time it took for him to read the passage. The test administrator provided these time measurements by writing on the chalkboard the elapsed time in ten-second intervals. The instructions did not urge the students to read at any particular speed but, if they stressed anything, it was that the student's reading was primarily to prepare him to take the completion test that followed. These tests were administered by the teachers regularly employed by the schools. These teachers were provided with appropriate training for this task, and their work was monitored by a member of the project staff. Rate of reading was defined as the number of words in the passage divided by the number of minutes required to read the passage.

Only three grade levels of students were used, grades 4, 7, and 10. Just 20 students were tested at each grade level, and so each student took eight booklets, one at each of the eight difficulty levels. The booklets were administered on successive days at a rate of one booklet each day. The students were drawn from the same school district as in Study III. The grade 4 and 7 students each represented intact classrooms. It proved necessary to select the grade 10 students by asking for volunteers and then selecting from among these volunteers a group having a distribution of achievement performances similar to the distribution of the total group at that grade level.

Score Adjustments

The adjustments made to these scores were, in most respects,

identical to those made for Study III. The chief contrasts arose from the fact that repeated measures were taken on the students tested.

Results

These data were used in various ways to verify the results of Study III; however, only the regression of rate on cloze is of interest here. Plots of these data showed clearly that there was a difference between the regressions for each grade level. The students at the higher grades appeared to show more rapid rates of reading than the students at lower grade levels. Unfortunately, however, there was not sufficient overlap in the distributions of cloze scores to provide useful estimates to be made of the grade level effects on the regression curves. Consequently, only cloze was used in the regression equation presented here and subsequently used in the model. This equation is shown in Table 10.

TABLE 10

Regression Equation for Rate of Reading on Cloze Scores

(R = .44, S.E. = 55.76)

$$\text{Words per Minute} = 3.3796C + .1535C^{4} - .00004606C^{8} + 77.291$$

The curve described by this regression is shown in Figure 21. This regression appears to be primarily linear. The hook in the upper extreme occurred principally because the column means in that range showed a sharp decline. However, it must be kept in mind that this curve represents a composite of the curves for each of the three grade levels. The plots for each of those grade levels were such that one would expect the shapes of the curve for each grade level, when they are eventually determined, to somewhat resemble those of the post-reading regression shown in Figure 8. That is, they will probably fan out from the low cloze scores to the high cloze scores and perhaps level off somewhat in the upper range of cloze scores. The point at which the curves level off seems likely to occur at lower levels of cloze performance for students at lower grade levels.

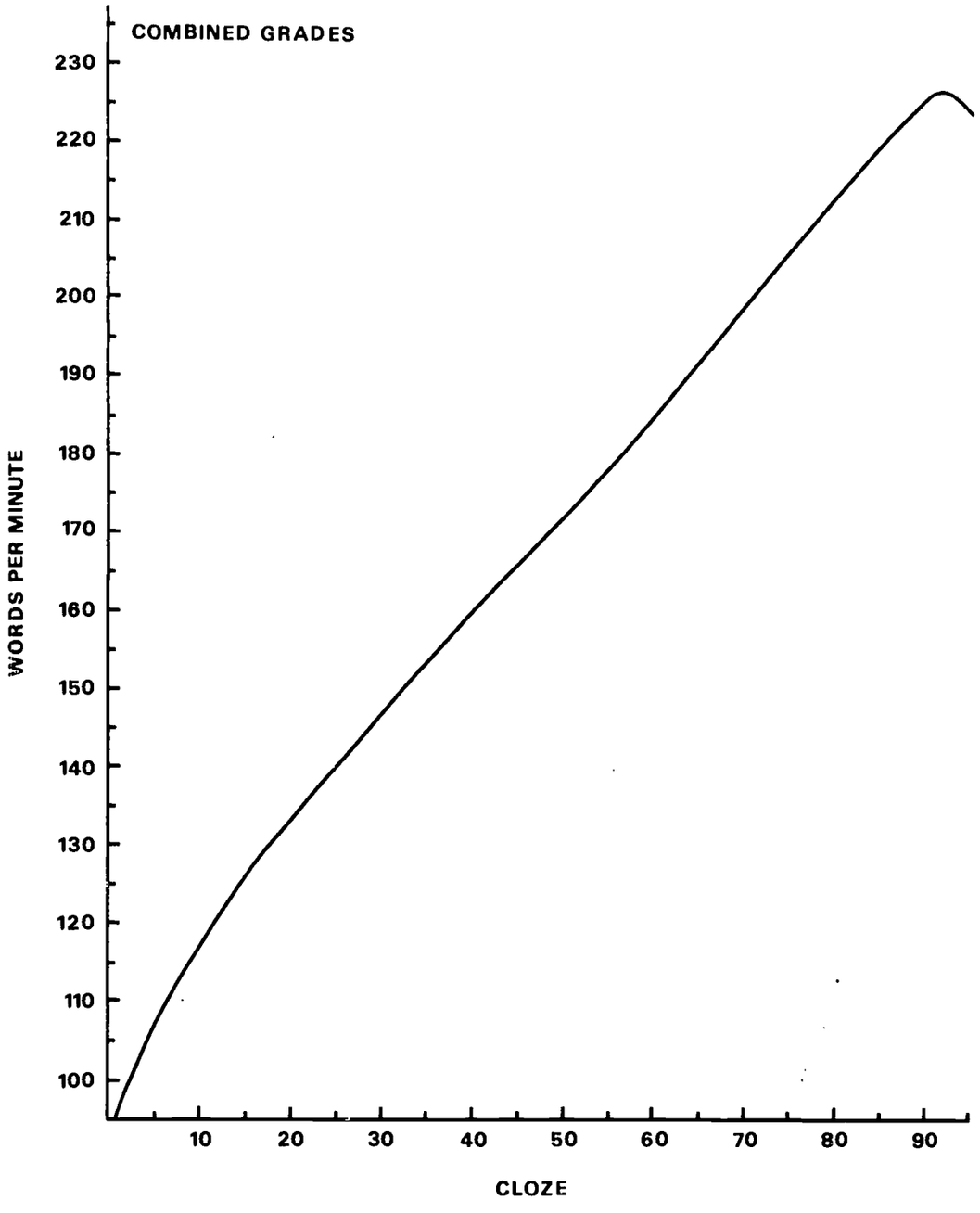


FIGURE 21. Regression of words read per minute on cloze scores.

Remarks

It should go without saying that this study provided only a preliminary estimate of the regression rate of reading on cloze scores. However, since a fair amount of care was taken in selecting the students for this study, it seemed more reasonable to use even this regression in the criterion selection model than to use no regression at all to represent rate of reading. It is true that the regression is averaged across what may turn out to be some rather large grade level effects and that that average is probably somewhat biased due to the fact that there were not equal numbers of students at each grade level attaining each level of cloze score. However, including even this regression in the criterion selection model seems to be preferable to omitting measures of proficiency altogether and thereby completely omitting a major category in the taxonomy of benefits accrued from reading a passage.

STUDY V

Purpose

This study determined three sets of relative weights for the behaviors included in the criterion selection model, one set for each of the three major uses to which instructional materials are put--textbook, reference, and voluntary reading. The method used was to have teachers rate the relative importance of the behaviors. The criterion selection model requires that each behavior be weighted according to its relative importance before the values representing each behavior are summed to determine the value associated with a particular level of performance on a cloze test. The mean ratings by teachers were obtained in this study to serve as an approximation to these weights. This study also attempted to determine if the uses for which materials were considered or the grade level and subject matter taught by the teacher doing the ratings affected the weights.

Rationale

It should be clear from the outset of this study that weights obtained by the methods employed here can serve only as approximations of the weights that must ultimately be obtained. In the first section of this report it was explained that the weight that should be assigned to a behavior in the model was actually determined as a sum of the products based on the degree to which each behavior contributed to each cost and benefit of instruction and the value society set upon each of the costs and benefits. However, the costs and benefits of instruction have never been determined explicitly for any area of instruction, nor do we have estimations of how each behavior in the model is related to those costs and benefits. Consequently, it seems impractical at the present time to estimate the weights for the passage performance criterion model using direct procedures.

However, even though the costs, benefits, and relative values of these outcomes of instruction may not have been made explicit, it seems likely that to some extent they implicitly govern much of the instruction in a school, being embodied in such instruments as the curriculum guides, textbooks, training of teachers, standardized tests and other indices used to assess the effectiveness of instruction, and so on. Moreover, since teachers are fairly intimately acquainted with these instruments and generally maintain communication with the parents of the students, it seems likely that the teachers may develop a fairly accurate sense of what the costs and benefits are, of how they are valued, and of how they relate to the behaviors included in the model, at least to the extent that these matters relate to the teachers' own responsibilities. Thus, teacher ratings of the relative importance

of the behaviors in the model ought to provide a fairly good approximation to the relative weights that might be obtained through more formal methods.

However, a caveat should be inserted at this point to deny that weightings obtained by teachers' ratings, even if they were a perfectly accurate estimate of what the weights are in actual practice, could replace weights derived through rational policy-making procedures. Perhaps the most important purpose of creating a model for policy-making processes is to determine what these weight values ought to be in order to represent most effectively the values of society and in order to attain society's goals most efficiently. At best, weights based on teachers' ratings can only represent a normative measure of what those weights actually are.

Procedures

Materials: Each teacher participating in this study was given a booklet that consisted of a cover page and three scales, each scale appearing on a separate page. The first scale had them rate the relative importance for textbook reading of information gain, rate of reading, subject matter preference, style preference, difficulty preference, and willingness-to-study preference. The next two scales had them rate the same behaviors but for reference and then voluntary reading. Two versions of information gain were rated in these scales, gain measured by completion tests and gain measured by multiple-choice tests. The ratings for gain measured by multiple-choice tests were later ignored when it was decided not to use the measures from Study I to select a set of performance criterion scores. Figure 22 shows an example of one of the rating scales used. Notice that the descriptions of behaviors are stated in operational terms and that the description of the preference behaviors adhere fairly closely to the wordings of the items in the preference ratings administered to students.

Administration of the Scales: The scales were administered to groups of teachers drawn from the same schools as the students in Studies III and IV. As the booklets were passed out, the teachers were asked to fill in identification data on the cover sheet, information that included the grade level and, for teachers at grades 7 to 12, the subject matter they taught. When a teacher was responsible for teaching more than one grade level or subject, he was asked to identify the category in which he dealt with the most students and to confine his responses to just that category. The instructions began with a brief and general description of these studies and then instructions were given on how to fill out the scales. The teachers were told to read each description of a behavior, to select the one that was most important and place its letter under the number 10, to select the least important one and place its letter under the number 1, and then to rate

LEAST IMPORTANT	1	2	3	4	5	6	7	8	9	10	MOST IMPORTANT

A. INFORMATION GAIN ON COMPLETION TESTS: How important is it for a student to answer many more completion questions about the content of a textbook selection after he has read it than he could answer before he read it.

B. INFORMATION GAIN ON MULTIPLE CHOICE TESTS: How important is it for a student to answer many more multiple choice questions about the content of a textbook selection after he has read it than he could answer before he read it.

C. RATE OF READING: How important is it for a student to read a textbook rapidly.

D. SUBJECT MATTER INTEREST: How important is it for a student to say that he likes very much to learn about the things talked about in his textbook regardless of whether he learns them by talking to others, watching T.V., reading a book, or listening to a teacher.

E. PREFERENCE FOR WRITING STYLE: How important is it for a student to say that he thinks a textbook is very interestingly written (not so much what the author says, but how he says it).

F. SUITABILITY OF DIFFICULTY LEVEL: How important is it for a student to say that a textbook is at about the right level of difficulty for him to read, neither too easy nor too hard.

G. WILLINGNESS TO STUDY: How important is it for a student to say that he would like very much or would be very willing to study a certain book as the regular textbook for one of his classes.

FIGURE 22. Scale used by teachers to rate the relative importance of the behaviors for coping with the reading tasks involved in their instruction.

the remaining behaviors relative to these two behaviors. They were asked to avoid ties whenever it seemed reasonable to do so but were told that the extra lines under each number were provided to be used in case of ties.

Although about 107 teachers were tested, 101 booklets were obtained in which the teachers had followed directions. The numbers of teachers obtained at each grade level varied from 6 to 14. Only regular classroom teachers were used at the elementary grade levels and only mathematics, science, social studies, and English teachers were used in grades 7 to 12.

Results

The major analysis was to determine whether teachers made consistent discriminations among the behaviors being rated and whether the uses to which materials are put, the subject matter taught by the teacher, or the grade level taught by the teacher influenced the teachers' ratings. Two analyses were made of the variances. The first was made just of the ratings by teachers in the upper grade levels, grades 7 to 12, where information was available on the subject matter taught by the teacher. Grade, subject matter, use, and behavior were treated as completely crossed and fixed factors, and teacher was treated as random and as nested within grade and subject. Two teachers were randomly selected from the appropriate groups to represent the teachers in each cell. In the other analysis teachers from all grade levels were used, but the subject matter taught was disregarded and six randomly selected teachers from the appropriate groups appeared within each cell. The results are shown in Table 11.

The main effects of grade, subject taught, and use were not significant. However, this cannot be interpreted as meaning that those factors had no influence on the teachers' ratings, since all teachers were required to distribute their ratings over the full range of the scale, from 1 to 10. A significant effect on one of these factors would therefore have to be interpreted as indicating that the teachers within strata on those factors tended to cluster their responses differently on the scales. Since these effects were not significant, it can be said that the teachers at all levels on these factors tended to distribute their ratings in about the same ways on the scales. The large and significant effects due to the behavior rated, however, show that there was substantial agreement among the teachers on their respective ratings of different behaviors.

The effects of grade level, subject taught, and use would, however, show up in their interactions with each other and with the behaviors being rated. Only the subject matter-by-behavior and the use-by-behavior interactions were significant. The significant subject matter-by-behavior

TABLE 11

Analyses of the Variances in Teacher Ratings

Source of Variation	Upper Grades Only			All Grades		
	d.f.	M.S.	F	d.f.	M.S.	F
Grade, G	5	9.23	2.40	9	3.78	F<1
Subject Taught, S	3	7.72	2.01			
Use, U	2	.23	F<1	2	.65	F<1
Behavior, B	6	645.18	56.17**	6	868.50	75.16**
GS	15	7.55	1.96			
GU	10	1.47	F<1	18	1.01	F<1
SU	6	1.51	F<1			
GB	30	13.41	1.17	54	11.30	F<1
SB	18	26.92	2.34*			
UB	12	57.02	18.37**	12	64.72	20.70**
Teacher, T(GS)	24	3.85				
T(G)				50	5.65	
GSU	30	1.28	F<1			
GSB	90	10.49	F<1			
GUB	60	3.53	F<1	108	3.36	1.08
SUB	36	3.53	F<1			
TU(GS)	48	1.64				
TU(G)				100	1.59	
TB(GS)	144	11.49				
TB(G)				300	11.56	
GSUB	180	3.35	1.08			
TUB(GS)	288	3.10				
TUB(G)				600	3.13	

*p<.05
**p<.01

interaction was fairly small but nevertheless important, for it indicates that teachers in different subject matter areas rate the relative importance of variables somewhat differently and that it may be necessary to expand the criterion selection model to take the subject matter of the materials into account. The use-by-behavior interaction was also significant and relatively large. Since use and behavior were already treated individually by the model, no expansion of the model was required by this result; but it did require that a separate weight be used for each behavior under each use condition.

The mean ratings assigned by the teachers are shown in Table 12. These means were based on all 101 cases. The behavior-by-use interaction appears clearly in some of these means. Information gain was weighted as less important in reference and voluntary reading than in textbook reading, while the ratings of the importance of style and willingness-to-study preferences showed the reverse pattern. The pattern seems to suggest that teachers believe that it is relatively less important for students to understand materials whose study they do not supervise directly, but that students' feelings about those materials become increasingly important as the teacher can exercise less direct supervision over their study. While this pattern does not seem altogether the most desirable one from the point of view of an adult's need to pursue independent study, it seems to reflect realistically the relative importance of the behaviors in the various kinds of reading tasks the student must perform in order to cope with the teacher's testing and grading practices. Consequently, in view of the type of performance criterion scores to be identified in these studies, that is, criterion scores for judging the suitability of materials to be used in an instructional setting, this pattern of ratings has a considerable amount of common-sense plausibility.

Discussion

Weights for the behaviors in the criterion selection model were determined from teachers' ratings of the relative importance of those behaviors. A fairly plausible argument was advanced to support the proposition that these weights would probably reflect fairly accurately the values actually placed on those behaviors, and it was shown that teachers exhibited a high degree of agreement on their ratings. However, it must be stressed that 101 schoolteachers could conceivably be individually and unanimously wrong in this matter. When these weights are eventually determined rationally, the weight assigned to each behavior in the model at that time might, for example, be set as proportional to the sum of the products of the values society places on each of the cost and benefit outcomes of instruction and the degrees to which the behavior contributes to each outcome. It is anything but certain that teachers can subjectively estimate these relationships without some degree of bias.

TABLE 12

Weights Assigned to the Behaviors

Behavior	Use		
	Textbook	Reference	Voluntary
Information Gain, Completion	5.53	4.89	2.73
Information Gain, Multiple Choice	5.10	4.79	2.74
Rate	2.74	2.85	3.53
Subject Matter Preference	8.68	8.51	8.69
Style Preference	4.78	4.55	7.26
Difficulty Preference	7.19	7.57	7.26
Willingness-To-Study Preference	7.87	8.29	9.12

However, the weights obtained seemed fairly plausible in view of the high degree of agreement among teachers and in view of the conformity of the results to a common-sense analysis of the instructional situation in which the behaviors must operate. The most important result from the point of view of the structure of the criterion selection model was the fact that the subject matter taught by the teacher had some influence on his ratings of the importance of the behaviors. This result suggests that the next efforts to elaborate the model should give high priority to determining how the subject matter of the materials influences the various regressions included in the model.

IDENTIFICATION OF PASSAGE PERFORMANCE CRITERION SCORE

This section demonstrates the criterion selection model in its present form by identifying a set of passage performance criterion scores. The chief objective of this series of studies was to demonstrate that rational procedures could be used to identify performance criterion scores for tests that represent a domain of content. And the more specific object was to obtain reasonably acceptable passage criterion scores for use in readability assessment procedures. Both objectives seemed attainable at this point. The model was sufficiently elaborated that it included behaviors that might have large effects on the levels at which the performance criterion scores are set, and the data gathered in Studies III, IV, and V had sufficient quality and scope to warrant placing some faith in performance criterion scores based on them. However, it should probably be stressed once more that these performance criterion scores should be regarded not as finished products, but rather as first approximations to the criterion scores that will eventually be settled on, as the model and its instrumentation are improved to represent more faithfully a consensus of how this decision ought to be made.

A total of 30 performance criterion scores, a set of three for each of the ten grade levels, was calculated. This was made necessary by the fact that both the grade level of the student and the use for which materials were considered had fairly large effects on the shapes of the regression curves. Moreover, use also had an effect on the relative weights teachers assigned to the behaviors in the model. Consequently, it was necessary to calculate a criterion score corresponding to each of the three uses of materials for each of the ten grade levels of students.

Method of Calculation

The criterion scores were calculated first for materials used for textbook reading. These calculations were performed using the regression equations for information gain, subject matter preference, and rate of reading plus the textbook reading equations for style, difficulty, and willingness-to-study preferences. These were the equations numbered 3, 4, 5, 8, and 11 in Table 9 and the rate of reading equation from Table 10.

Each of these equations was first corrected for arbitrary scale effects by dividing it by the standard deviation of the scores on the dependent variable. These standard deviations are shown in Table 13. The equation was then weighted by multiplying it by the relative weight teachers had assigned the behavior represented by that equation. These weights were presented in the left-hand column of Table 12.

TABLE 13

Standard Deviations of the Dependent Variables

Dependent Variable	S.D.
Information Gain	102.860
Subject Matter Preference	1.930
Rate of Reading	61.826
Style Preference	
Textbook Reading	1.230
Reference Reading	1.225
Voluntary Reading	1.317
Difficulty Preference	
Textbook Reading	.822
Reference Reading	.790
Voluntary Reading	.807
Willingness-To-Study Preference	
Textbook Reading	1.310
Reference Reading	1.280
Voluntary Reading	1.360

Finally, the six equations were summed to form one very long equation. Since all of the behaviors represented in this model are stated as benefits, or positively valued outcomes, the signs joining the equations were all positive. For example, suppose that rate of reading had been expressed in terms of minutes per word instead of words per minute. It would then have represented a measure of the cost of reading a passage in terms of the time and effort associated with the task and, therefore, would have been treated as a negative benefit by subtracting its effects in the equation. The same reversal of values could be performed on any of the other scales as well. However, the accounting problems are usually reduced if all variables are stated as positive benefits, as number correct instead of number wrong or as interest instead of boredom, for example.

The calculations then proceeded by setting the grade level value of this extended equation at 3 and solving successively for every integer value of the cloze scores falling within the observed range. These numbers defined the curve that represented the overall value a student in grade 3 would normally be expected to receive from reading a passage, if he attained any one of the possible cloze scores. The grade level value was then incremented to 4 and the process repeated to obtain the curve for grade 4 students, and the whole process was continued until curves for all ten grades had been obtained. The point where each of these curves reached a maximum value defined the performance criterion for that grade level of student on materials used for textbook reading.

Calculations of exactly the same type were used to determine the criterion scores for reference and voluntary reading. The equations used to calculate the reference reading criterion scores were the rate of reading equation from Table 10 and equations 3, 4, 6, 9, and 12 from Table 9. The teacher-assigned weights for each of these equations were taken from the center column in Table 12. The equations used to calculate the voluntary reading criterion scores were the rate equation from Table 10 and equations 3, 4, 7, 10, and 13 from Table 9. The teacher-assigned weights for each of these equations were taken from the right-hand column of Table 12.

Illustration of the Calculation Procedure: The effects of these computations can be understood best, perhaps, by examining the illustration represented by Figure 23. This set of curves was obtained using equations in which grade level was omitted, and so it represents roughly the results that would have been obtained from solving the equation for textbook reading by students in grade 7.5. These curves were obtained by solving first the weighted information gain equation for each of the cloze scores in the observed range of cloze scores and subtracting out of each value the distance between the cloze axis and the lowest point on this curve. For example, this obtained a value of about 5.5 for a cloze score of 30 percent. Thus, the height of this curve at a given level of cloze performance shows the weighted information gain value a

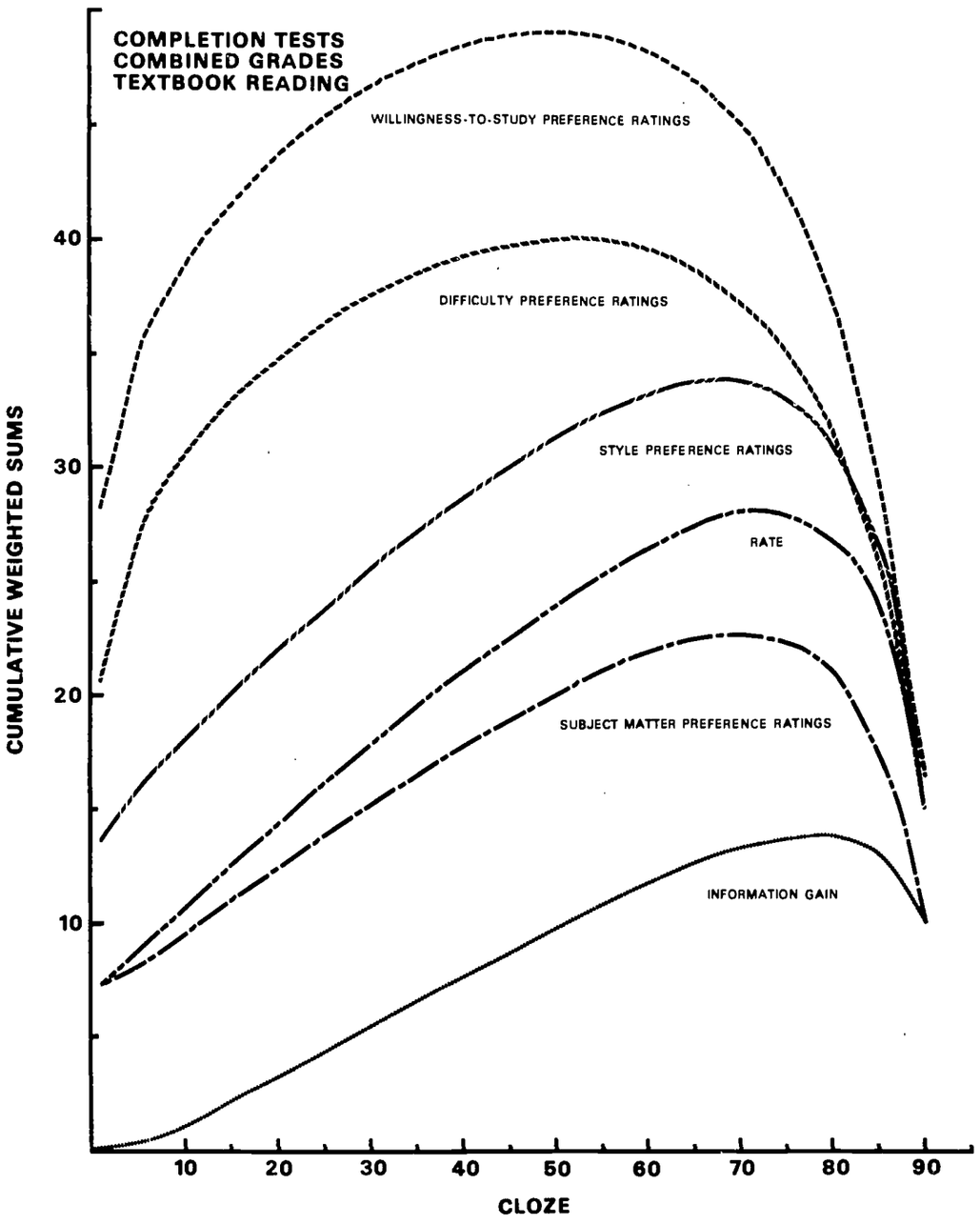


FIGURE 23. Illustration of how the operations specified in the model arrive at a criterion score. The model successively sums the results of each weighted regression equation to arrive at the top curve, the cumulative weighted sum. The cloze score vertically below the highest point on this top line is defined as the passage performance criterion for the particular grade level of students and use of materials represented by that curve.

student is ordinarily expected to obtain from reading a passage at that level of cloze performance. The exact numeric size of these scores has no particular meaning. Rather, the meaning of the value is derived from its size in relation to the other values on that curve. Hence with respect to information gain, a cloze score of 30 percent is worth a little more than a third as much as a cloze score of 70 percent. This line could be said, then, to represent a component of the curve that finally resulted as the top line of the graph.

The next curve, labeled subject matter preference ratings, was obtained by performing the same operations with the weighted subject matter preference regression equation, the next component of the weighted sum, and then adding the value obtained for each cloze score to the value obtained previously for that cloze score from the information gain equation. For example, the values obtained at a cloze score of 30 percent were roughly 5.5 from the information gain equation and roughly 10 from the subject matter preference equation, and so the cumulative height of the curve when it contains both information gain and subject matter preference is the sum of these numbers, 15.5, at a cloze score of 30 percent. The top line of the graph was reached when this process was repeated for each successive equation. The height of a point on this top line represents the weighted and cumulatively summed values associated with the cloze score below that point on the curve. The cloze score directly below the highest point in the top curve represents the level of performance at which these weighted and summed values are at a maximum. This is the cloze score that the model defines as the passage performance criterion score for the particular grade level of student and materials represented by the curve. Henceforth, the results of these calculations will be presented only for the top line in each set of curves.

Criterion Scores: Figures 24, 25, and 26 show the results when grade level was taken into account. Each curve in one of these figures represents the top line from a calculation of the weighted sum for a single grade level of students. The graphs for textbook, reference, and voluntary reading are presented in that order. The only difference between the calculation used to arrive at these curves and the one used to obtain the top line presented in Figure 23 was that the lowest point on each component curve above the base line was not subtracted out in the present calculations. Consequently, the differences in the vertical heights of the curves for different grade levels stems from the fact that similar differences in height were observed in the regression equations that make up the components of these sums. Leaving this information in these curves affects only the height of the top line of the weighted sums, not its shape or the point at which the curve reaches its maximum.

Table 14 presents two kinds of information. The left-hand column of numbers represents the passage performance criterion scores identified by the criterion selection model. Each of these numbers represents the point at which one of these curves reached its maximum value. To be explicit, it is the point in the observed range of each grade level's

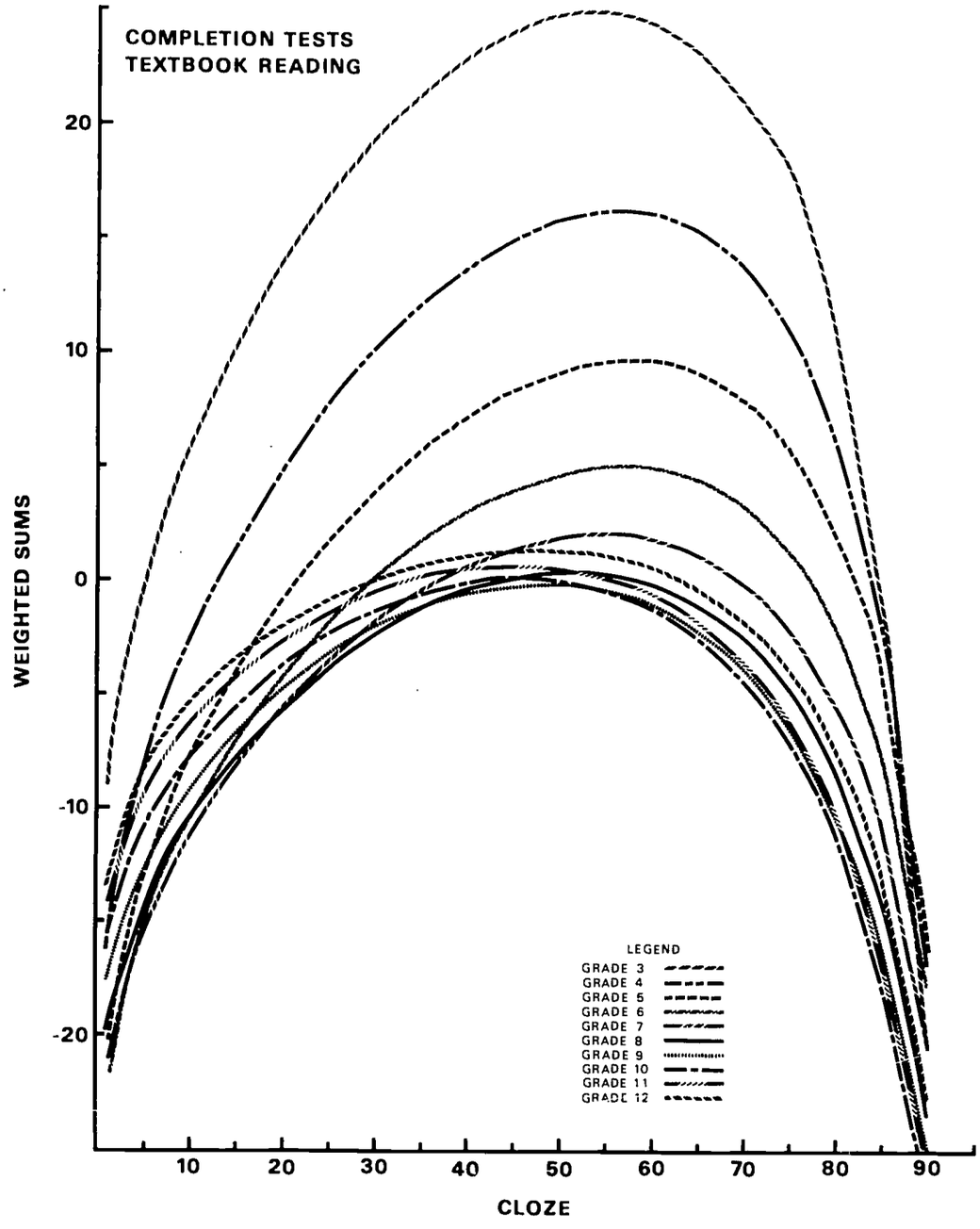


FIGURE 24. Weighted sums of the regression equations used to identify criterion scores for textbook reading. The point at which each curve peaks represents the passage performance criterion score for that grade level of students for textbook reading purposes.

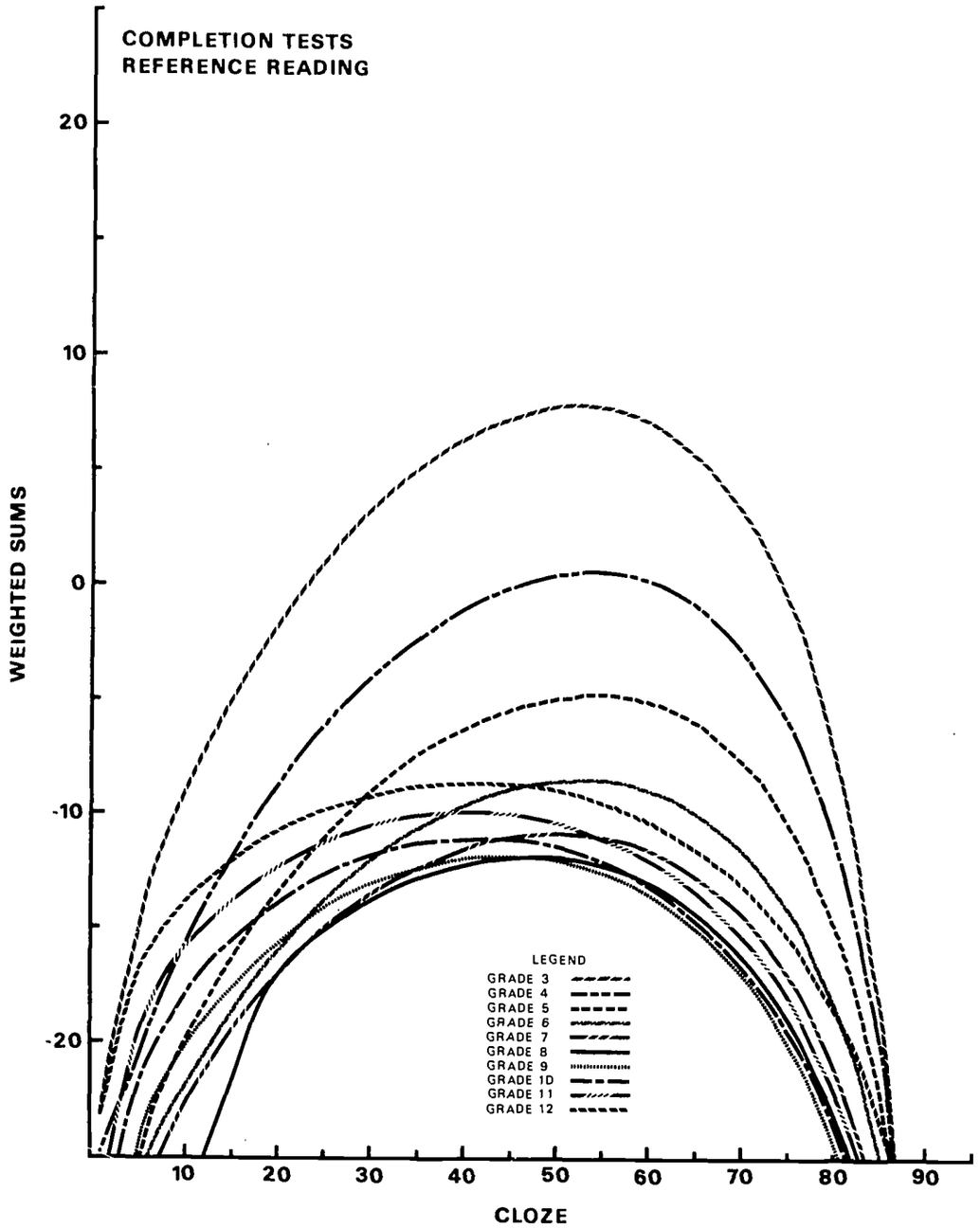


FIGURE 25. Weighted sums of the regression equations used to identify criterion scores for reference reading. The point at which each curve peaks represents the passage perforamnce criterion score for that grade level for reference reading purposes.

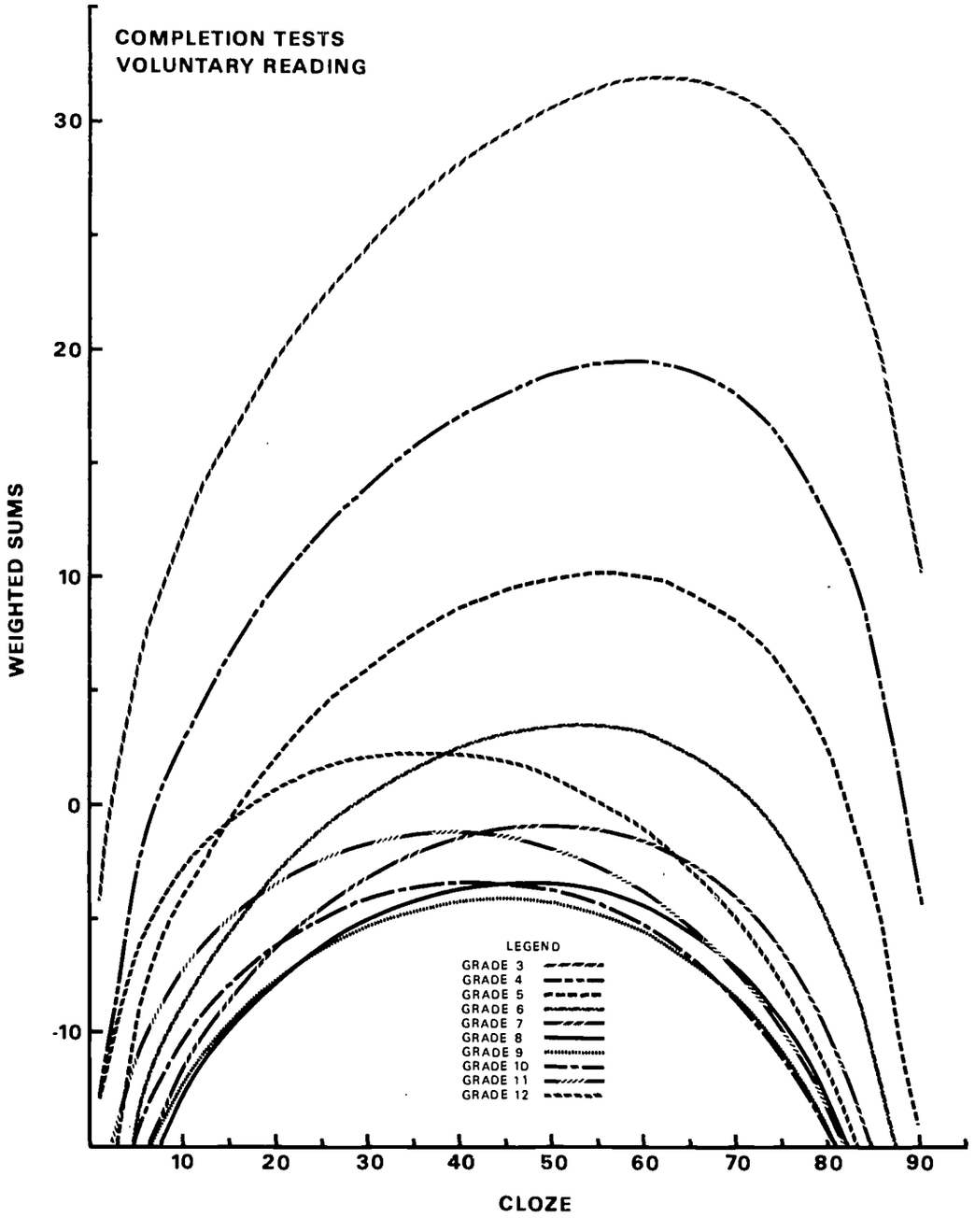


FIGURE 26. Weighted sums of the regression equations used to identify criterion scores for voluntary reading. The point at which each curve peaks represents the passage performance criterion score for that grade level for voluntary reading purposes.

TABLE 14

Passage Performance Criterion Scores and the
Efficiency Rates They Produce on the Dependent Behaviors

Criterion	Cloze Criterion Score	Dependent Behaviors					
		Information Gain	Rate of Reading	Subject Matter	Style	Difficulty	Willingness-To-Study
				Grade 3			
Textbook	54[a]	81	59	100	99	55	100
Reference	52	78	57	100	98	47	100
Voluntary	62	90	68	97	99	1	100
				Grade 4			
Textbook	56	81	61	100	99	56	100
Reference	54	78	59	100	97	56	99
Voluntary	58	83	63	100	99	28	100
				Grade 5			
Textbook	57	80	62	99	98	58	100
Reference	54	76	59	98	97	63	99
Voluntary	55	77	60	98	99	47	100
				Grade 6			
Textbook	57	78	62	98	97	61	100
Reference	52	71	57	96	98	68	99
Voluntary	52	71	57	96	99	59	100

TABLE 14 (Continued)

Grade 7							
Textbook	54	73	59	99	97	67	100
Reference	50	67	55	97	98	70	99
Voluntary	50	67	55	97	99	65	100
Grade 8							
Textbook	51	69	56	100	97	71	100
Reference	47	63	52	99	98	71	98
Voluntary	47	63	52	99	99	69	100
Grade 9							
Textbook	48	64	52	100	97	74	99
Reference	44	58	48	100	98	71	98
Voluntary	45	59	49	100	100	71	100
Grade 10							
Textbook	46	61	50	100	95	73	98
Reference	42	55	46	100	97	69	98
Voluntary	42	55	46	100	100	73	100
Grade 11							
Textbook	45	60	49	100	92	70	98
Reference	41	50	43	99	95	69	98
Voluntary	39	50	43	99	100	75	100
Grade 12							
Textbook	48	67	52	98	79	60	97
Reference	41	54	45	95	84	62	96
Voluntary	36	44	40	91	100	78	99

[a]Cloze criterion scores are expressed as percentage of items answered correctly.

curve of weighted sums where the first derivative of the equation for weighted sums, differentiated with respect to cloze, equalled zero. For example, note the curve for students in grade 3 on textbook reading in Figure 24. This curve reached its maximum value at a cloze score of approximately 54, the number shown as the criterion score for that grade level and use in Table 14.

Efficiency Rates: The numbers to the right of each criterion score in Table 14 are presented as an aid in judging the plausibility of these criterion scores. Each number shows how high a component curve, the information gain curve for example, was at the point where the criterion score was set. These numbers were calculated by subtracting the height of the lowest point of the curve from the height of the highest point of the curve to obtain the height span of the curve, next subtracting the height at the lowest point of the curve from the height of the curve at the point where the passage performance criterion was set to obtain the net height of the curve at the criterion score, then dividing the net height by the height span to determine what proportion of its height the curve had reached at the point where the criterion was set, and finally multiplying this proportion by 100 to express this quantity as a percentage. Thus, for example, the information gain curve for grade 3 students in textbook reading was at 81 percent of its height span at the point where the criterion score was set. Or, stated somewhat differently, grade 3 students who obtain a cloze score of 54 percent on a passage are likely to gain 81 percent of the maximum amount of information they would be likely to gain at any level of cloze performance. And, finally, this number could be regarded as an efficiency rate of a sort. Thus, the criterion score of 54 percent may be regarded as producing an efficiency rate of 81 percent on information gain, 59 percent on rate of reading, and so on. The discussion that follows will employ the terminology of efficiency rates.

Evaluation of the Model

The important question that must be dealt with at this point is whether the criterion scores produced by the model in its current stage of development are sufficiently valid to be of use in practice. This is not to raise the fundamental question of whether the model faithfully and fully represents the logic and evidence by which these criterion scores ultimately ought to be identified. That question was dealt with when the model itself was discussed. Rather, it is merely to raise the pragmatic question of whether the model has produced sensible results in its current stage of development, results that are sufficiently sensible to be used in practice as reasonable approximations to the criterion scores that will ultimately be identified.

Three of the results seem central to deciding this issue. First, it appears that the subject matter, style, and willingness-to-study

preference ratings are the dominant variables in determining where the criterion score is set by the model. Second, the model dictates that as students reach higher grade levels they should be given more difficult materials. And third, the model generally dictates that students be given easier materials for use in textbook reading than for reference and voluntary reading.

Domination by Preference Ratings: It seems clear from the efficiency rates shown in Table 14 that the dominant effects in deciding where the criterion scores would be set were produced by the subject matter, style, and willingness-to-study preference rating regressions. Moreover, it is also possible to construct a fairly plausible argument that this dominance may have represented a bias that implicitly added the effects of a single preference variable into the model at least two times. Consequently, these criterion scores may reflect a bias that gives undue weight to the preference variables in the model.

The efficiency rates in Table 14 show clearly that these three preference ratings had the dominant effect on setting the level of the criterion scores. First, note that the criterion scores showed a total range of variability of about 20 percentage points and that the rate, information gain, and difficulty preference efficiency rates were also fairly variable. However, the efficiency rates of the subject matter, style, and willingness-to-study preference ratings were not only uniform, regardless of the cloze level at which the criterion scores were set, but also were always very near 100 percent, an effect that could occur because these three sets of curves had roughly the same shapes and because they were each assigned fairly high weights by teachers. Thus, the criterion scores were nearly always set at the point where the regression curves for these three preference ratings reached their maximum values. And this effect largely overrode the effects of the remaining variables in the model.

The dominance of these three preference regressions would be legitimate if each represented an independent behavior. However, there was some evidence that they are highly redundant, possibly measuring a single behavior. If so, they would bias the model by, in effect, including the same behavior repeatedly. The evidence for this speculation was that these preference scales were highly correlated and their regressions were similar in form. However, this is not to state the trivial and obvious proposition that the ratings were merely correlated with each other. Correlations could be induced by fortuitous correlations in instruction and other equally irrelevant considerations such as the confounding of novelty of content with passage difficulty. Rather, it may be speculated that some sort of hierarchic dependency may exist among the ratings such that the willingness-to-study ratings, for example, might arise from a process that involves the student's reactions to the subject matter, style, and difficulty of the passage, summarizing them in a single response measure. If such dependencies exist, the effect would be to bias the model by adding the same response

to the model a number of times under different aliases. The nature of this proposition can be sharpened, perhaps, by pointing out that even though rate and information gain are fairly highly correlated, the same assault cannot be made on their independence. The weight of evidence indicates that they are hierarchically independent, since they can be manipulated fairly independently through the instruction of a student and the arrangement of the reading task.

Finally, it must be noted that even if a dependency of this sort occurred among the preference scales, and even though that dependency would work primarily to reduce the efficiency rate of information gain, the model still did not produce absurd results. The information gain measure maintained an efficiency rate of more than 50 percent in nearly every case.

Effects of Grade Level on Criterion Scores: It can be seen from Table 14 that the model specified lower criterion scores for older students. To some extent, this effect was also a result of the dominance of the preference ratings in the model. Figures 9 to 20 show that the information gain curves rose more sharply at lower cloze scores and that the preference curves tended to peak at progressively lower cloze scores as the students got older. This would, in part, account for the fact that criterion scores for older students were set at lower cloze scores. However, the role played by the preference ratings in producing this effect can be seen from the fact that the efficiency rates of the preference ratings remain high and even increase with grade level while the efficiency rates for information gain decline fairly steadily. Consequently, if the model in its present form gives too much weight to the preference ratings, the effect of correcting that bias would be to reduce somewhat the differences between the criterion scores assigned to different grade levels of students, but not to eliminate that difference.

Easier Materials for Textbook Reading: The third result of major interest was the fact that the model specified that students should receive easier materials for use in their supervised study, for textbook reading, than they should get for their independent study, for reference and voluntary reading. What makes this a result of major interest is the fact that it runs exactly counter to the practices recommended by teacher trainers in the area of reading.

The practice usually recommended by authors of teacher training materials in the area of reading (Betts, 1946; Bond and Tinker, 1967; and Harris, 1962) is to give students harder materials for their supervised study than for their unsupervised or independent study. The rationale generally given for this traditional practice is that a teacher's instruction in giving a reading assignment provides the student with help on the vocabulary and other matters that might cause him difficulty as he reads. Moreover, the class discussion that supposedly follows a reading assignment is thought to clear up any remaining problems caused

by the difficulty of the materials. Reading that the student conducts independently, on the other hand, must proceed without this assistance, so the materials used for this purpose should be easier and thereby assure that the student can learn the contents of those materials. Also, these authorities cite what was heretofore taken to be a fact, that the easier materials are for a student, the better he likes to read them. And they reason from this that a student is more likely to actually perform independent reading tasks if the materials are sufficiently easy for him. Thus, this rationale assumes that the teacher introduces the materials and that class discussion functions to fill any gaps in the student's knowledge left by his inability to deal effectively with the content; and that students prefer to read easy materials.

This rationale seems to rest on false premises at nearly every step. The most obvious one is the assumption that the easier materials are for a student the more he will like them. The regressions involving the preference ratings refute this proposition with a fair degree of finality. Students prefer materials at a level of difficulty that they perceive as appropriate for them, and their preference ratings drop at an increasingly sharp rate as the materials become either harder or easier. Consequently, the traditional practice could not, under any circumstance, be justified in these terms.

This contradiction of the traditional recommendations could not have arisen from a use-by-difficulty interaction effect on the preference scales. Although this interaction was significant on the style and difficulty scales, as shown in Table 6, it could not have produced this effect since it acted in opposite directions on the ratings for reference and voluntary reading. The form of the interaction was to depress the ratings of hard materials for voluntary reading relative to the ratings for textbook reading and to elevate the ratings for the hard materials for reference reading relative to the ratings of textbook reading. Had this interaction been the source of this contradiction, the contradiction would have occurred only on one of the two uses, reference or voluntary reading. The criterion scores for voluntary reading would have conformed to traditional recommendations and would have been set at higher cloze scores than the criterion scores for textbook reading, while the criterion scores for reference reading would have been lower than the criterion scores for textbook reading, contradicting the traditional recommendations. However, these effects would have been barely noticeable since the interaction was quite small numerically.

Rather, the effect seems to have arisen primarily from the teachers' ratings. Table 12 shows that the teachers tended to rate information gain as much more important in textbook reading than in reference and voluntary reading. The effect of these weightings was to draw the criterion score for textbook reading toward the higher cloze scores where the information gain curves were at their highest levels. At the

same time, the teachers rated style preference and willingness to study as more important for materials used for reference and voluntary reading purposes. Since these regression curves tended to reach their maxima at lower cloze scores than the information gain curves, the effect of their increased weightings was to draw the criterion scores for reference and voluntary reading toward the lower cloze score.

Thus, this result seems to have occurred primarily as a consequence of a difference of opinion between reading experts and teachers on what behaviors are most important in the various uses of materials. Possibly the teachers regard the textbook as a primary and more or less self-contained instrument for imparting knowledge. They may not want to spend much time in reteaching the contents of a textbook and otherwise remedying its deficiencies. Consequently, they would place a high premium on the students' ability to independently gain the information in the textbook. The reading expert, on the other hand, may not be aware of the realities of classroom practice or he may regard this remedial function as an acceptable use of the teacher's and the student's time in class discussion, and perhaps regard a degree of incomprehensibility in textbooks as acceptable or at least inevitable. In any case the traditional practice of setting the criterion for instructional reading lower is clearly incorrect since a performance criterion should reflect the actual conditions of instruction, not merely someone's notion about what those conditions ought to be.

Summary: In summary, then, there is a possibility that the model in its present form may contain a bias that might be due to a large degree of overlap in the processes measured by the preference scales. The chief evidence for this bias is the fact that the ratings were highly correlated and that many of the preference regression curves were similar in form, evidence that is suggestive but ambiguous. However, if the preference ratings are redundant, then the effects would generally be to set the criterion scores at lower cloze scores than should be the case, and to induce a systematic decrease in criterion scores as the students got older. On the other hand, the model did seem to conform better to the realities of the way materials actually function in instruction than the criterion scores recommended by writers of teacher training materials in the area of reading instruction. Moreover, at no time did the model produce criterion scores that were absurd. Even information gain, which would have been the chief victim of the bias that might have occurred in the model, maintained a moderate efficiency rate in even the most extreme cases.

Evaluation of the Criterion Scores

Passage performance criterion scores can be put to a number of uses described elsewhere in some detail (Bormuth, 1969a and 1970a). And the criterion scores presented here are probably much superior to any

performance criterion presently in use, even though these new criterion scores represent only the crude first approximations to those ultimately sought as passage performance criteria. However, they should not be used by practitioners and researchers without considerable caution, since they contain both systematic and random error. Consequently, their strengths and weaknesses should be carefully studied. The following discussion briefly commences this analysis by pointing out the major sources of systematic and random error.

Systematic Error: At least five sources of systematic error influence these criterion scores, each producing a bias of some form. The first was just discussed in some detail. This was the bias that may have been induced by the possibility that hierarchic dependencies existed among the preference rating responses. The curves representing the preference ratings generally peaked at lower cloze scores than rate and information gain. Consequently, if the preference variables were over-weighted, the effect would be to draw the criterion scores to levels that are lower than is warranted.

A second source of bias, the omission of major outcomes of reading a set of materials, was discussed briefly as a part of the critique of the model. Some of these omissions were fairly obvious. For example, the information gain measures should ideally contain items testing information gained by inferential processes as broadly defined, the rate of reading measure should be augmented with more direct measures of response proficiency and latency, and the cognitive measures should include transfer and retention. However, other omissions may not be quite so obvious, particularly to the psychologists who are most likely to participate in this work. One of these is the economic or cost factor, which is so far represented only very indirectly in the model. If materials cannot be understood by the students who have to read them, education funds are being wasted and the student's value as a productive member of society is being reduced and possibly being forced into the negative column.

Similarly, the psychosocial factor must be taken into account. It is probably not implausible to speculate that forcing students continuously to use materials that are inappropriate for them can produce effects on how they think of themselves in relation to society. Regard the student who is continually confronted with materials that are beyond his grasp and who is seldom able to acquire the knowledge and experience his fellow students find easy access to in those materials. In effect, this person has been ostracized from an important set of activities in his society, and so it would not be improbable for him to develop defenses against the society, defenses that could be costly both to him and to society. The high rates of illiteracy among juvenile delinquents might, in part, index this effect.

Thus, the criterion scores offered here are based on an incomplete model. At the present time it is not clear how the biases represented by these omissions from the model might influence the criterion scores. The psychosocial factors might tend to draw the criterion scores toward higher levels of cloze performance, while the cost factors, including the publishers' costs of adjusting materials to suit students having an array of reading abilities, might tend to draw the criterion scores to lower levels of cloze performance. In any case, the criterion scores presented here should be recognized for what they are, performance levels identified by a model that requires further development.

The teacher weightings may have also constituted an important source of systematic bias. The rationale for obtaining the weights from teacher ratings necessarily assumed that teachers have a knowledge, or at least an unverbalized sense, of what the outcomes of instruction ought to be, of how society values these outcomes, and of the degrees to which the behaviors in the model contribute to the attainment of those outcomes. On the one hand, it is decidedly risky to challenge judgments merely because they are based on clinical experience. The human mind seems to be capable of subjectively integrating highly complex phenomena and of making accurate estimates of this sort. However, it is equally well known that superstitions and worthless old-wives' remedies derive from the same processes. Consequently, the type of weights used in this model must be regarded as a temporary expedient and as a potential source of bias that has unknown effects on the criterion scores.

A fourth possible source of bias may have arisen in the sample of students used to determine the regressions. Nearly all were white, middle-class students living in suburban communities. While the relative homogeneity of this group may have had little effect on the regressions with respect only to cloze scores, the students' grade level also influenced the levels at which the criterion scores were set. It is difficult at this point to determine just what this grade level dimension represented and, therefore, how well the results can be generalized to other populations of students. On the one hand, it may have represented a general level of academic achievement, a variable that might be more effectively operationalized by averaging the students' scores on a wide-range battery of achievement tests. If this were the case, then the grade level dimension of these performance criterion scores would not generalize well to the remainder of the population, for these students were decidedly superior in achievement to average students. On the other hand this dimension may have represented some more deep-seated aspect of the student's development, an aspect that would be less subject to environmental influences than school achievement. For example, it should be noted that the grade level effects generally diminished greatly in most of the regressions at about the time puberty was reached, suggesting that these effects may have arisen as a consequence of biological development factors rather than merely from educational factors. If this were the case, then the grade level dimension might generalize to other

populations fairly well. However, at this time the grade level dimension must be regarded with suspicion, as a potential source of bias, in applying the criterion scores to students who were not similar to those used to obtain the regressions.

A fifth source of bias in these regressions may have arisen from a failure to take into account all of the factors that influence the parameters of the regressions. Two such sources were identified in the course of these studies. One was the difficulty dimension of the passages and the other was the subject matter of the material. The passage difficulty dimension was ignored in these studies because it was unclear what this dimension might represent and because it did not account for a large amount of variance. The subject matter dimension, however, seems at least at the present time to represent a more detailed breakdown of the uses to which materials are put. In any case these and as yet unidentified factors may have an influence on the levels at which the criterion scores should be set. Consequently, this should be regarded as a source of some systematic bias in the criterion scores presented here.

Random Error: The estimates of the criterion scores were also influenced by several sources of sampling error; that is, by errors in the measurements of the students' abilities and preferences, errors in sampling students, errors in sampling passages, and errors in drawing samples of items from the passages. To some degree, each source of sampling error increased the error involved in estimating the true values of the criterion scores, that is, the criterion scores that would be found if the study were repeated an infinite number of times and the criterion scores found at each replication were averaged. However, it is not clear at this time how one would set confidence bounds around these criterion scores to represent their accuracy.

Advantages of These Criterion Scores: After this recital of the bias and error that may influence the accuracy of these criterion scores, it may seem incongruous to state that they are suitable for application in practice. However, this is a relative rather than an absolute judgment. That is, the criterion scores presented here are better by several orders of magnitude than those presently in use.

The criterion scores presently in use require the teacher to write a set of comprehension test questions over a passage drawn from materials, to administer them as a post-reading test to students, and then to use the materials for textbook reading just with those students who answer at least 75 percent of the questions correctly, and for reference and voluntary reading just with those students who answer at least 90 percent of the questions. Earlier discussions in this report dealt with the arbitrary nature of these criterion scores and with the fallacies in the rationale supplied for them, and so only a few observations need to be added.

One of the most serious flaws in the traditional procedures arises from the fact that the teacher writes the test questions. Few teachers are trained in the skills required to construct test questions that can meet even reasonably loose standards of replicability. Consequently, the difficulties of these tests probably vary greatly from one teacher to another and from one time to another for the same teacher, possibly depending on accidental variables such as how generous the teacher happens to feel at the time he writes or scores the test. Consequently, it is highly uncertain what a score of 75 or 90 percent might represent on these tests. The performance criterion scores identified here, however, utilize tests made by the cloze procedure, and this procedure economically produces tests that have uniform relationships to the passages from which they are made. Consequently, using the passage performance scores presented here has at least the advantage of uniformity of meaning from passage from passage.

The major strength in the criterion scores presented here lies in the model and the data by which they were derived. That is, a criterion score derives its validity from what it produces for its user. The traditional criterion scores produce essentially unknown outcomes for their users, whereas the present criterion scores are supplied with a considerable amount of data to describe many of their effects. However, it must be emphasized that this is not to claim that the criterion scores presented here were good in any absolute sense. Rather, it is claimed here that they are far better than the criterion scores currently in use. The earlier discussions in this report should have made it abundantly clear that much work remains before criterion scores that are good in an absolute sense can be supplied to educators. The identification of performance criterion scores for criterion reference tests is analogous in many ways to assigning norms to norm reference tests. And there is no reason to believe that developing the theory and technology for identifying performance criterion scores will be any less complex or less energy-consuming than the development of the theory and technology for assigning norms.

REFERENCES

Allen, J. E. The right to read--Target for the 70's. Address delivered before the National Associations of State Boards of Education Convention, Los Angeles, September 23, 1969.

Berlyne, D. E. Conflict, arousal, and curiosity. New York: McGraw-Hill, 1960.

Betts, E. A. Foundations of reading instruction. New York: American Book, 1946.

Block, J. H. The effects of various levels of performance on selected cognitive, affective, and time variables. Unpublished doctoral dissertation, University of Chicago, 1970.

Bloom, B. S. Individual differences in school achievement: A vanishing point? Paper presented at the American Educational Research Association Convention, New York, February, 1971.

Bloom, B. S. (Ed.) Taxonomy of educational objectives, handbook I: Cognitive domain. New York: David McKay, 1956.

Bond, G. L., and Tinker, M. A. Reading difficulties: Their diagnosis and correction. New York: Appleton-Century-Crofts, 1967.

Bormuth, J. R. Reading literacy: Its definition and measurement. Symposium organized by the Reading Committee of the National Academy of Education, 1970a.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970b.

Bormuth, J. R. Development of readability analyses. (USOE Tech. Rep. No. 7-0052.) U.S. Office of Education, 1969a.

Bormuth, J. R. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969, 4, 358-365b.

Bormuth, J. R. Cloze readability criterion reference scores. Journal of Educational Measurement, 1968, 5, 189-196.

Bormuth, J. R. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading, 1967, 10, 291-299a.

Bormuth, J. R. The implications and use of cloze procedure in the evaluation of instructional programs. Los Angeles: University of California Center for the Study of Evaluation of Instructional Programs, Occasional Report No. 3, 1967b.

Bormuth, J. R. Readability: A new approach. Reading Research Quarterly, 1966, 1, 79-132.

Bormuth, J. R. Validities of grammatical and semantic classifications of cloze test scores. Proceedings of the International Reading Association, 1965, 10, 283-286.

Bormuth, J. R. Cloze tests as a measure of comprehension ability and readability. (Doctoral dissertation, Indiana University, 1962.) Ann Arbor, Michigan: University Microfilms, 1963. Number 63, 2586.

Bormuth, J. R., Manning, J., Carr, J., and Pearson, D. Children's comprehension of between- and within-sentence syntactic structures. Journal of Educational Psychology, 1970, 61, 349-357.

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Carroll, J. B. Vectors of prose style. In Style in language. Thomas A. Sebeok (Ed.). Cambridge, Mass.: John Wiley and Sons, 1960.

Coleman, E. B., and Miller, G. R. A measure of information gained during prose learning. Reading Research Quarterly, 1968, 3, 369-386.

Cornell, F. G., Lindvall, C. M., and Sciupe, J. L. An exploratory measurement of individualities of schools and classrooms. University of Illinois Bulletin, 1963, 50 (Whole No. 75).

Cronbach, L. J., and Furby, L. How we should measure "change": Or should we? Psychological Bulletin, 1970, 74, 68-80.

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

Dember, W. N. A new look at motivation. American Scientist, 1965, 53, 409-427.

Dember, W. N., and Earl, R. W. Analysis of exploratory, manipulatory, and curiosity behavior. Psychological Review, 1957, 64, 91-96.

Fillmore, C. J. The case for case. In Universals in linguistic theory. Emmon Bach and Robert T. Harms (Eds.). New York: Holt, Rinehart and Winston, 1968.

Flesch, R. F. Estimating the comprehension difficulty of magazine articles. Journal of General Psychology, 1943, 28, 63-80.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Guttman, L., and Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. Educational and Psychological Measurement, 1967, 27, 569-580.

Harris, A. J. Effective teaching of reading. New York: David MacKay, 1962.

Harris, C. W. (Ed.). Problems in measuring change. Madison: University of Wisconsin Press, 1967.

Hively, W., II, Patterson, H. L., and Page, S. H. A "universe-defined" system of arithmetic tests. Journal of Educational Measurement, 1968, 5, 275-290.

Johnson, P. E. On communication of concepts in science. Journal of Educational Psychology, 1969, 60, 32-40.

Johnson, T. J. A conceptual learning model and instructional implications. St. Ann, Missouri: Central Midwestern Regional Educational Laboratory, 1971.

Kamman, R. Verbal complexity and preferences in poetry. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 536-540.

Klare, G. R., Sinaiko, H. W., and Stolurow, L. M. The cloze procedure: A convenient readability test for training materials and translations. (Institute for Defense Analyses, Paper No. P-660.) Arlington, Virginia: Institute for Defense Analyses, Science and Technology Division, 1971.

Kretch, D., and Crutchfield, R. S. Elements of psychology. New York: Knopf, 1958.

Lorge, I. Readability formulae--An evaluation. Elementary English, 1949, 26, 86-95.

Lorge, I. Predicting reading difficulty of selections for children. Elementary English Review, 1939, 16, 229-233.

Martin, H. C., and Ohmann, R. M. The logic and rhetoric of exposition. New York: Holt, Rinehart and Winston, 1965.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. The measurement of meaning. Urbana: University of Illinois Press, 1967.

Potter, T. C. A taxonomy of cloze research, Part I: Readability and reading comprehension. (Southwest Regional Laboratory for Educational Research and Development Tech. Rep. No. 11) Los Angeles, 1968.

Rankin, E. F., and Culhane, J. W. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading, 1969, 13, 193-198.

Rankin, E. J. Cloze Procedure--A survey of research. Yearbook of the Southwest Reading Conference, 1965, 14, 133-148.

Rothkopf, E. Z. Learning from written materials: An exploration of the control of inspection behavior by test-like events. American Educational Research Journal, 1966, 3, 241-249.

Schlesinger, I. M., and Weiser, Z. A facet design for tests of reading comprehension. Reading Research Quarterly, 1970, 5, 566-580.

Singer, H. Factors involved in general reading ability in the content areas. Paper presented at the National Reading Conference, Atlanta, 1969.

Taylor, W. L. Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 415-433.

Thorndike, E. L., and Lorge, I. The teacher's word book of 30,000 words. New York: Bureau of Publications, Teachers College, Columbia University, 1944.

Weaver, W. W., and Kingston, A. J. A factor analysis of the cloze procedure and other measures of reading and language ability. Journal of Communication, 1963, 13, 252-261.

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

APPENDIX A

PASSAGES

SCHOOL HELPERS

The school nurse is a good helper. She helps the children remember good health rules. She weighs and measures the children in the nurse's room. She helps boys and girls if they get hurt. She helps make school a happy and a safe place.

The librarian likes to have boys and girls visit the school library. She helps them find good books to read. She helps them find books to answer their questions. She shows boys and girls how to take care of the books.

In the school kitchen are the cafeteria helpers. They wear clean, white clothes. Every day they cook good food for hungry boys and girls. They have many dishes to wash. Before they go home, they leave the kitchen nice and clean.

The crossing guard holds up his big stop sign. He tells the cars when to stop. He tells the cars when to go. He tells the children when to cross the street. He helps the children remember to walk carefully across the street.

The school bus driver is a careful driver. He always remembers the safety rules. He brings the boys and girls to school in a big yellow bus. He makes sure they get to school on time. After school he takes them back to their homes again.

The music teacher visits many schools. She is a special school helper. She helps boys and girls learn many songs. They learn funny songs. They learn pretty songs. They learn fast songs. They learn slow songs.

TIME: _____

FIXING UP AN OLD HOUSE

"I want to know what they are going to do. Let us go over and ask them," said Mary.

The children went over to the builders and asked them what they were going to do. The builders said they were going to make the old house into a two-family house. One family would live on the ground floor. The other family would live on the top floor.

The man who used to live there had asked that the house be made into a two-family house. Then people would pay him to let them live there.

Every day the children looked at the house. From time to time they stopped to see what the builders were doing. There were three men at work all the time. One man saw that the work followed the plan. He saw to it that everything went well.

One day the children watched the men put a door in the back wall. They did not go too near the house or get in the way.

When it rained, the men worked in the house. The children could not go in, but they looked in the windows. Then they could see what the men were doing. They watched the men put up a wall to make two rooms out of one. All the men worked on the wall together.

One day other men were working in the house. They put wires in the new wall. They put in a bell for the back door. Then they went away.

TIME: _____

PROBABILITY

There is a branch of mathematics called probability which tells us how often we can expect certain things to happen.

One example of probability is what happens when we toss a coin. A coin has two sides -- heads and tails. If you toss the coin once, it will either land heads or tails. There are only two ways it can land.

Suppose you toss a coin two times. It could land on heads both times. It could land on tails both times. It could land heads the first time and tails the second time. It could land tails the first time and heads the second time. There are four different ways the coin could land with two tosses. It is more likely that it will land with one head and one tail than with two heads or two tails.

If you toss the coin ten times, you have a better chance of getting five heads and five tails than any other combination.

If you toss a coin one hundred times it is likely that you will get between 45 and 55 heads. It is not very likely that you will get 90 heads.

If you toss a coin twenty times it is not likely that all twenty will land heads.

Words such as "likely," "good chance," "not likely" are used to show the probability of things happening in the future.

Try tossing a coin many times to test some ideas of probability. Keep a record or a tally of each toss.

Another example of probability is the roll of a die. A die has six sides. There are six different ways it can land. Try rolling the die and keeping a record of the way it lands.

Tally marks are useful for collecting information on the rolls of the die.

TIME:_____

CHANGE OF THE MOON

The moon does not always look the same. The moon seems to change. The moon seems to get bigger and bigger each night. After the moon is big, it seems to get smaller and smaller. The moon does not really change. It does not get bigger or smaller. It just seems to change.

The sun lights a part of the moon. You see the part that the sun lights. Sometimes you see all of the lighted part. Sometimes you see some of the lighted part. Sometimes you see just a little of the lighted part. Sometimes you cannot see any of the lighted part. You cannot see the moon all the time. You cannot see the moon when it is on the other side of the earth.

The earth turns around. You go with it. The moon seems to come up in the east because you are turning from west to east. The moon seems to go down in the west because you are turning away from the moon. You are always turning from west to east. Because you are turning, the moon seems to come up and go down in the sky.

The sun is in the sky. The sun is a star. Stars make light and heat. Stars are heavenly bodies.

The moon and earth are in the sky. The moon and earth are heavenly bodies. The moon and earth do not make light and heat. The moon and earth get light and heat from the sun.

Mars is a heavenly body. Mars is something like the earth. It does not make light or heat. Mars gets light from the sun. You can see Mars from the earth. Mars is so far away that it looks like a star. You cannot always see Mars.

TIME: _____

A-4

EVAPORATION

Wash the blackboard. Watch it dry. The water goes into the air. When water goes into the air it evaporates.

Tie a damp cloth to one end of a stick. Tie a bottle to the other end. Put water in the bottle until the stick is level. Watch the stick for a few minutes. It does not stay level.

Water goes into the air when it evaporates. It changes into water vapor. You cannot see water vapor, but it is in the air all around you.

Cut a hole in the bottom of a cardboard box. Hold the box against a cold window and blow into the hole. Take away the box. Look at the window. There is dew on the cold glass. Water vapor in your breath changed into drops of water. When water vapor changes into drops of water it condenses.

Fill a can with water and ice. Fill another can with warm water. Let them both stand a few minutes. Now look at the cans. Touch the outside of each one. There is water on the outside of one can. The water must have come from the air near the can. Water that comes on cold things from the air is called dew.

Put hot water in a glass. Put a cool, dry glass upside down over it. Water in the lower glass evaporates. The water vapor rises into the upper glass. The water vapor is cooled by the cool glass. The water vapor condenses into drops of water.

TIME: _____

SMALL STREAM

This is a pool of water. The water is cool. The water is clear. The water is still. The still pool is like a mirror. You can see the sky in the water. You can see white clouds in the water. You can see your face in the water. Water flows from the pool. The water makes a path. Water comes from other pools. The path gets deeper and deeper. The water makes a brook.

Sometimes the water moves slowly. Sometimes the water moves fast. The water jumps and leaps over rocks. Sometimes the brook goes through a field. Sometimes the brook goes through a forest. The water in the brook carries many things. The water carries sticks and soil, grass and sand. Twigs and seeds and leaves fall into the brook. Animals come to the brook. They drink the cool water. Some animals live near the brook. They swim in the water. Animals stir up the water. They stir up sticks and twigs. They stir up grass and soil, seeds and leaves. Now the water is muddy. Now the brook is not clean. What will clean the brook? Fish clean the brook. They eat bugs and flies. Rocks, pebbles, and sand clean the brook. They make screens in the brook. Screens hold things back. A fence is a screen that holds things back. A net is a screen that holds things back. Screens in windows hold things back. In the brook, rocks make a big screen. Rocks have spaces between them. Water goes through the spaces. Branches and leaves cannot go through the spaces. The big screen holds them back.

TIME: _____

MISCHIEVOUS LITTLE PEAR

There was once a Chinese boy called Little Pear. He lived with his father, his mother, and his two sisters in a small house at the edge of a village in China. All around the village were flat fields of cabbages and beans and onions, and far away on one side was a great highway that led to the city, and far away on the other side was a river. Little Pear's mother used to say to him, "You may run and play out of doors, but do not go too near the river. You might fall in." Sometimes, though, Little Pear would disobey his mother. He loved to stand on the high bank and look down on the swift muddy river and the ships sailing down it toward the sea. He would hold very tight to a huge willow tree with both hands and think, "I can never fall in -- not if I am very careful, like this."

Little Pear was a very mischievous child. His sisters said he was naughty. His father said he was naughty and would cry, "Ay-ah! What a bad boy you are!" But his mother said, "He is very little; when he gets bigger he will be good; you wait and see. It doesn't matter if he is naughty now, sometimes!" And they all loved Little Pear very much.

Little Pear was five years old. He had a round, solemn face with eyes like black apple seeds. He didn't look mischievous at all. His head was shaved, except for one round spot just over his forehead where the hair was allowed to grow and was braided into a little pigtail tied with bright-colored string.

TIME:_____

CHANGING WATER POWER TO ELECTRICITY

Have you ever wondered where electricity comes from? If you say water power, you will be about half right. You will be about half right, because almost half of the electricity used in the United States comes from water power. Where does this water power come from? How is the water power changed into electricity?

In some places, there are big waterfalls. The force of falling water is used to make electricity.

In some places, where there are no waterfalls, people have dammed up streams and rivers. Then they force the water to fall from a great height.

Hydroelectric plants change water power into electricity. One big hydroelectric plant can supply electricity to many cities, even though they may be some distance from the plant.

The word hydroelectric makes you think of waterfalls but at most hydroelectric stations you seldom see falling water. You see water spilling over the walls of dams built for power plants only when the water level gets too high. Sometimes the narrow streams flowing down mountains are used for hydroelectric power. When there are stations at the foot of such streams, you never see the falling water.

How does water get into the hydroelectric plant? You cannot see it flow into the station, because it enters through huge iron pipes. These pipes are called penstocks.

At the bottom of each penstock there is an enormous wheel called a turbine. This wheel has curved blades in it. The falling water dashes against the blades and whirls the turbine. The turbine turns a long steel rod. This spinning rod is connected to the generator and makes the generator go. The generator makes electricity.

TIME: _____

CAT'S EYES

All of the Big Cats, as well as the lesser ones, have wonderful eyes. They can see clearly even on a dark night. This is because of the way they are made. There is a sort of window in each eye. This window is called the pupil. It is black and is placed in the center of the colored part of the eye.

The pupil lets light come in and reach a kind of mirror at the back of each eye. These mirrors reflect everything that is in front of the eyes. Right away a special nerve carries these reflected pictures to the brain. Then the brain sends a quick signal to all parts of the body. This signal may be to attack, hide, be careful, or perhaps run away. Whatever the signal may be, the cat automatically follows it.

These reflections will be poor unless the right amount of light reaches the eye mirrors. So the pupils open and close on their own. On sunny days the light is very strong, so the pupil of the cat eye is only a narrow slit. When there is less light in the evening, the pupils become rounded and much larger. At night they seem to cover the whole eye. This means that they are getting in all the light possible to make good reflections.

These pupil changes are made quickly, and without the cat knowing it. You can prove this with a pet cat. Some bright day, when his pupils are just narrow slits, carry him into a dark closet. Keep him there for a minute or so. Then turn on the light and look at his eyes. Their pupils will be round and wide open. But if you keep the light shining, the pupils will soon become small again.

TIME:_____

REFINING CRUDE OIL

Petroleum which comes directly from the ground is called crude oil. Crude means unprepared. Crude oil is petroleum which is not ready to be used in machines or by people. Crude oil must be refined before it can be used. When crude oil is refined, it is changed and made ready for use. At a refinery, crude oil is changed into different products and made ready for use.

Crude oil is made of many particles called molecules. Molecules are too small to be seen. Each molecule is made up of several smaller particles called atoms.

In a molecule of pure oil, there are several atoms called hydrogen. There are also several atoms called carbon. Molecules which contain these two kinds of atoms are called hydrocarbons. Almost all petroleum products are made of hydrocarbons.

When water is heated and boiled, part of the water goes into the air as vapor. The steam from a steam kettle is vapor going into the air.

When water vapor is cooled, it condenses. This means it changes back into water.

Crude oil can be boiled and changed into vapor. Crude oil vapor can be condensed and changed back into oil. When crude oil is changed into vapor and back into a liquid, it is distilled. This process is called fractionation.

Crude oil is made of different combinations of hydrocarbons. Each of these different combinations of hydrocarbons will change into vapor at a different temperature. Some of the kinds of oil will change to vapor with just a little heat. Some of the combinations must have a lot of heat to change into vapor.

TIME:_____

THE BEGINNING OF POST OFFICES

Whenever people want to send messages to each other, they can write letters. People write letters to tell some news. People write letters to sell things like furniture or cars. They write letters to ask for jobs. They write letters to give invitations.

There are so many different reasons for sending messages. Long ago, people had the same reasons for writing letters that we have today. But they did not have an easy way of sending the letters. It took many days, or even weeks, for a letter to go a short distance. A piece of mail to another part of our country had to carried in a slow boat or by a friend who was traveling.

In the early part of the last century, the Pony Express was set up to send mail from place to place. A messenger with a bag of mail would ride his horse as fast as he could until he reached a particular spot. Then he would stop and give the mail to another messenger on a fresh horse. Each stopping place was called a "post." That is why we call our mail buildings "post offices."

You have probably been to the post office in your neighborhood. Maybe you went by yourself or maybe you went with someone in your family to buy stamps or to mail a package. The largest post office in each city is usually downtown. It is called the main post office, while the smaller neighborhood ones are called branches. Larger cities need large buildings to handle the thousands of pieces of mail which go in and out of them every day.

TIME:_____

A-11

HOW NUMBERS ARE USEFUL

When people learned how to count they could tell each other many things.

Suppose a man went fishing and when he came home his wife said, "What did you catch?" If he said, "I caught 12 fish" she knew right away that he had this many. She didn't have to go down to his boat and see if he had caught a lot or just a few.

Then if the fisherman went into the next room to wash while she cooked supper and she shouted to him and said, "How big were they?"

What would he do? Come all the way in there and hold out his wet hands and say this big or this big? No. People had thought of a better way to tell than that and they called it Measuring.

Measuring is just telling certain sizes and weights and distances and parts of a day that everybody knows about and saying how many of those any new thing would be.

If the fisherman said to his wife, "One fish is as long as my foot," she would know how big it was right away. And if he said that another one was as long as 2 feet she'd know how big that was. Men used their feet so much to measure things that even today we make measuring sticks called rulers that are about as long as a big man's foot.

What do you think we do to measure something that isn't as long as a foot ruler? We make 12 marks on the ruler that are all the same distance apart from each other. Each one of those distances we call an inch. It is one inch from each mark to the next one. Now we can measure a little thing by telling how many inches long it is.

TIME:_____

MATTER IS MADE FROM ATOMS

When scientists talk of matter, they mean anything at all that has weight -- a rock, a human being, a book, a pail of sea water, or an automobile.

All matter is made up of tiny particles. These particles are so small that they cannot be seen with any microscope invented so far. They are called atoms.

As we look about, it must seem to us that there are thousands of different kinds of matter in the world. It may be surprising, then, to learn that the number of different kinds of atoms out of which all this variety is made is not very large. The kinds of atoms we know about, in fact, are only 102 in number. What's more, most of these 102 kinds of atoms are very rare. Some do not exist in nature at all; they are found only in the laboratories where scientists have made them. Only a dozen kinds of atoms are really what you might call common.

Atoms sometimes exist as separate particles without any special connection to other particles. Mostly, however, they form groups. Such groups of atoms are known as molecules. These groups stick together, more or less, as time passes. This behavior is similar, in some ways, to the behaviors of human beings.

Some people may live all by themselves, but most form part of a family. Although there are only two kinds of human beings, males and females, there are many different kinds of families. You can have just a man and his wife. You can have a widowed mother and three children, all girls. You can have an old married couple with a son, a daughter-in-law, and two grandsons. There are thousands of possible kinds of families.

TIME:_____

THE SHAPE OF THE EARTH

What would you think of the nature of this world if you had never read a geography book? You probably would never get the idea that the earth is spherical, like a huge ball, as you do not see it that way. You may go up on a high mountain and look around the great circuit of the horizon. It may look flat, perhaps hilly, but it does not give the impression that you are on a sphere.

No wonder people first thought the earth was a flat disk. They knew that if one went far enough in any direction, one was likely to hit the sea. So they made their disk float in the ocean. But earth is heavier than water, and the earth would not float unless hollow inside, like a buoy. Thus they thought that there was a huge cavern inside and that the spirits of their ancestors dwelt there.

Above the disk was sketched a rotating cover called the firmament, holding the sun, moon, and the planets. The stars were little holes in the firmament through which shone the glory of heaven above. What held the whole system together and what was beyond the waters they did not know. But do we know what is beyond the stars and galaxies?

This was a beautiful system and was taught in the schools of Babylon and Greece for a long time, but there were some disturbing facts. First of all, as they began to trade with India and sailed boldly through the Pillars of Hercules (Gibraltar), they found that they could go a great deal farther west and east than they could go north and south. So some geographers made the earth oblong instead of disk-shaped.

A-14 TIME:_____

MEASUREMENT

When we drive into a service station, an attendant pumps gasoline into the tank of our automobile. The pump measures the gasoline in GALLONS as it feeds the fuel into the tank. The amount of MONEY we pay for the gasoline depends on the number of gallons measured by the pump. We can find the total cost by multiplying the price of one gallon by the number of gallons we received.

Buying gasoline shows us some things that are true of all measurements. MEASUREMENTS ARE NUMBERS. We use UNITS, such as GALLONS, INCHES, and HOURS, to measure things. To measure, we must find HOW MANY units there are in what we are measuring.

We can make some measurements DIRECTLY. For example, we can measure a kite string with a foot ruler by holding it against the string. But suppose we want to measure the distance a ship has traveled. We cannot apply a unit, such as a mile, directly. So we must find the answer INDIRECTLY with arithmetic, using the speed of the ship and the length of time it has been traveling.

Most measurements involve reading some kind of scale. No matter how many subdivisions the scale has, the object being measured is likely to fall between two of them. This means that EVERY MEASUREMENT IS AN APPROXIMATION. A measurement may come very close. But it never matches the scale perfectly. For example, with the unaided eye, we cannot read an ordinary ruler that is GRADUATED, or marked, much more closely than within sixty-fourths of an inch.

TIME:_____

VACUUM TUBES

The miracle of modern radio and television all depends on a wonderful invention called a "vacuum tube."

You've heard about "electronics." "Electronics" means electricity flowing through a vacuum instead of through a wire.

A vacuum is, simply, nothing. You take a glass tube and pump all the air out of it. Then you seal it so that the air can't get back in. The "nothing" inside the tube is called a vacuum.

A vacuum tube looks something like an electric light bulb. Inside it are some wires, a square of metal called a "plate," and a fine metal screen. You pump all the air out of the tube and seal it so that no more air can get back in. Now the wires and the plate and the screen are in a vacuum.

It is very hard to make electricity jump off a wire outside a vacuum because the air pushes it back. But in a vacuum there is nothing to keep it on the wire. You can make it jump right off and fly around in the vacuum.

You do this by making the wire hot, just as you make an electric bulb filament hot. The heat makes tiny bits of electricity jump off the wire. These bits of electricity are called "electrons."

The square of metal called the "plate" is beside the hot wire. There is always some electricity in a metal. If you pull it out, the metal will attract any nearby electrons. So you pull electricity out of the plate and push it into the hot wire through the wires that connect them.

TIME:_____

PLANT SURVIVAL

Plants and animals are the living, growing things of the earth. Most animals move around, and many of them have sound apparatuses, for making noises. Most plants spend their whole lives silently in one place. Because they live so quietly, we sometimes forget that during their growing seasons they work hard all day long. For plants are just as alive as animals, and they have the same problems: finding and keeping a place to live, getting food, fighting animal enemies and plant rivals, and having young, so that new plants will grow each year.

Scientists, taking hourly moving pictures of plants, then running them off quickly, have shown how active plants really are -- always twisting their stems and turning their leaves toward the sun, stretching out longer, growing new buds, then flowers, and finally seeds. Watch a sunflower as the sun moves across the sky, and see for yourself what happens.

People often think of plants' leaves and flowers as only decorative, and of their fruits as solely for human beings and animals to eat. But for plants themselves, their green leaves, their many-colored flowers, their great variety of fruits -- all the things about them -- are strictly business, part of the job of keeping alive and providing for new plants.

Plants are fierce rivals. Anyone who has ever weeded a garden knows that often several different kinds of plants are fighting for quarters on the same piece of ground. Each plant has its own special equipment to help it in the hard business of living. Part of the fun of knowing plants is discovering each kind's way of getting along in the world.

TIME:_____

THE BILL OF RIGHTS

Those who opposed the Constitution had shown tha many honest men were worried that their liberties were not safe. A number of states approved the Constitution with the understanding that certain rights and freedoms would be added to the document as soon as possible.

The men at the Convention were just as much concerned about liberty as the men who were against the Constitution. But they felt that first things should come first. If there was no strong government, people would not be safe in their lives or property, and they would not be free. Congress could be trusted, they believed, to protect liberty.

Some members of the Constitutional Convention wanted a bill of rights. That is why the delegates put in provisions protecting persons from "bills of attainder" (laws punishing a person without a fair trial) and forbidding laws which declared acts to be crimes after they were committed. The constitution also says that no man can be kept in jail without trial except in case of rebellion or invasion. But many people felt that the Constitution did not go far enough in protecting the people's freedom. The feeling was widespread. The Founding Fathers saw that the people would have to be reassured.

The new Congress which met in 1789 took action right away. James Madison proposed twelve amendments. Of these, ten were approved by the states and make up what we call the Bill of Rights or the First Ten Amendments. These became part of the Constitution in 1791.

TIME:_____

DISCOVERY OF AUSTRALIA

Australia lay remote and unsuspected. And yet, as the centuries passed, scholars in Europe came to believe that the southern ocean was not all empty.

"There must be a continent in the southern hemisphere," they said. "There has to be -- to balance the weight of Europe and Asia in the north. Otherwise the earth would be lopsided."

The scholars even gave a Latin name to the continent -- Terra Australis Incognita, the Unknown South Land.

In the early 1600's the Dutch, who had come as traders to the Spice Islands, finally discovered Australia. To them it was an extremely disappointing land. They found no gold, no spices, not so much as a single fruit tree. What was the good of a land like that?

An Englishman named Dampier heartily agreed with the Dutch in their opinion of New Holland, as the continent came to be called. He reported it to be the most barren spot on the globe. Its inhabitants, he said, were "the miserablest people in the world." "The Hottentots," he said, "though a nasty people, are gentlemen compared with these, who, setting aside their human shape, differ but little from the brutes."

But was New Holland the same as the Unknown South Land?

King George III of England, along with a lot of other people, did not believe it. No, somewhere in the South Seas lay another continent, a much better continent doubtless. King George wanted it found. Englishmen should take possession of the land in his name before somebody else laid claim to it.

TIME: _____

LIGHTNING

Lightning is a spectacular demonstration of the presence of energy at work in nature. When lightning occurs, its energy is changed into other forms of energy. Light can be seen in the flash. Heat from lightning sometimes starts fires. Mechanical energy is observed in the breaking of trees and other objects which have been struck. But the tremendous energy of lightning is largely uncontrolled and destructive. It is not in a form which can readily be used.

In his study of the merry-go-round of energy in nature, man has always been fascinated by lightning and has tried to understand it better and perhaps control it.

Benjamin Franklin was one of the first persons to study lightning carefully. He surmised that lightning was a discharge of electrical energy, perhaps a kind of static electricity or electricity in a state of rest. Others before him had done experiments with static electricity. They had observed the sparks that jump from the fur of certain animals when rubbed. They had noticed the sparks which jump from a person to a metal object after walking on a thick rug; they had noticed many examples of electrical energy jumping from one object to another.

Much was known about static electricity even before Franklin. It was known, for example, that certain objects charged with static electricity attract some objects and push others away. A hard rubber comb rubbed with fur will attract a glass rod rubbed with silk. On a small scale, then, static electricity can exert pushes and pulls.

TIME: _____

STATE LEGISLATURES

The legislature in New York and many other states is different from Congress when it comes to political parties controlling voting. In Congress members often -- although not always -- pay more attention to what a bill tries to do than they do to whether its sponsor is a Democrat or a Republican. In a state legislature, members more frequently vote on important measures strictly according to political party, with Republicans supporting their own bills and voting against Democratic ones, and vice versa.

This does not mean, however, that Democratic governors always get all the new laws they want from Democratic legislatures, or that Republican governors always have an easy time with Republican legislatures. They don't. Just as in Congress, controversies within the parties divide city dwellers from country people, liberals from conservatives, and the laws often come out a compromise.

Altogether, the government of a state is very much like the government of the country, only on a smaller scale. Each state governor has much the same type of power the President has, but only within the borders of his state.

A governor is in charge of all the departments that the state operates. He appoints the commissioners who run them. He decides how much money to ask the legislature to grant for each of them. If they run well, he gets credit. If they run badly, he's blamed.

The governor decides what new laws he thinks the state needs. He has his lawyers draw them up in bill form and send them to the legislature. Members propose many other bills that they draft themselves or get from businessmen, government officials or citizens' groups back home.

TIME:_____

PROPOSAL FOR TRANSPORTATION DEPARTMENT

President Johnson has asked Congress to set up a new Cabinet-level Department of Transportation. The department would coordinate and direct all federal programs dealing with transportation.

"The Founding Fathers rode by stage to Philadelphia to take part in the Constitutional Convention," said the President. "They could not have anticipated the immense complexity -- or the problems -- of transportation in our day."

The President mentioned traffic congestion and highway accidents as two major transportation problems.

The proposed transportation department would be one of the largest in the government. Nearly 100,000 employees from various existing government bureaus and agencies would come under it.

The department would take in the Federal Aviation Agency, the peacetime Coast Guard (now part of the Treasury Department), and the Bureau of Public Roads and the Maritime Administration (now in the Commerce Department). Safety functions performed by the Civil Aeronautics Board and the Interstate Commerce Commission would be taken over by the new department. The department would also assume some activities of the Army Corps of Engineers, the St. Lawrence Seaway Development Corporation, the Alaska Railroad, and the Great Lakes Pilotage Administration.

Under Mr. Johnson's proposal, a national safety board would be created in the department. The five board members would study safety standards and try to determine causes of land, sea, and air accidents.

The President also asked Congress to pass a traffic safety act. This act would be administered by the proposed transportation department. Under it, the department would ask auto and truck manufacturers to improve the safety of their vehicles.

TIME:_____

PROSPERITY AND WASTE

The people of the United States are in a sense becoming a nation on a tiger. They must learn to consume more and more or, they are warned, their magnificent economic machine may turn and devour them. They must be induced to step up their individual consumption higher and higher, whether they have any pressing need for the goods or not. Their ever-expanding economy demands it.

If modifications are forced upon the private-enterprise system of the United States in the future, it will be because that system did too good a job of filling many of the needs of the people. Defeat on such terms, we should all agree, would be saddening.

Man throughout recorded history has struggled -- often against appalling odds -- to cope with material scarcity. Today, there has been a massive break-through. The great challenge in the United States -- and soon in Western Europe -- is to cope with a threatened overabundance of the staples and amenities and frills of life. Conceivably, even the long-improverished and slower-starting Soviet Union may someday find itself trying to deal with an overflowing of goods. The United States, however, already is finding that the challenge of coping with its fabulous productivity is becoming a major national problem and is inspiring some ingenious responses and some disquieting changes.

Further -- and let's face it -- a good many Americans and Europeans have a pretty direct stake in the failure or success of businessmen in inducing us all to be more wasteful.

TIME:_____

A-23

THE ITALIAN

Gregariousness, curiosity and a fondness for communicating make Italians a universal people, the most universal in Europe. The provincial peasant with whom you share a railroad compartment is completely at home in your company. He may be part of a delegation traveling to a papal audience and it may be his first trip to Rome, but he is nonetheless a citizen of the world. At noontime he will offer you a chunk of his bread with slices of salami and a handful of shiny black olives. He will hand you his Chianti bottle with the crumbs from his lips still clinging to it. You are his traveling companion, and so he will want to know all about you. How much did your suit cost and where was the fabric woven? How long did it take you to cross the ocean? Have you ever been in Bridgeport, where he has relatives?

The peasant's love affair with the universe is his heritage from the peripatetic Romans who conquered the ancient world. The Caesars made themselves living symbols of universalism. Vergil and Livy gave form to the dream of a world community in their writings. A long history of invasions and occupations added to the original Roman and Etruscan blood the restless genes of Greeks, Goths, Normans, Spaniards and Moors. Through such a mixed strain, cosmopolitanism flows naturally. Pope Gregory the Great built a world-wide Christian community to which Thomas Aquinas gave the intellectual substance and Dante, humanistic poetry. Marco Polo's yearning for exotic civilizations turned him eastward, and the curiosity of Christopher Columbus and Amerigo Vespucci sent them west. Galileo explored the heavens.

TIME: _____

COMPOUND LOCI

Compound loci have previously been described in maize. The evidence for the duplicate nature of the locus usually comes from recombination experiments in which the two components of the complex are separated by rare crossovers. It is ordinarily difficult if not impossible to recognize the compound nature of a locus strictly on the basis of phenotype. If the action of the two genes in the complex are different enough to control different phenotypes, they are usually considered to be separate loci, even though they may have arisen by duplication of a single gene. The distinction between the single or compound form of a locus is easier in those cases where genetic differences in protein structure are analyzed, since single amino acid replacements can be detected electrophoretically if the replacement alters the net charge of the protein. A single cistron will specify a single protein form, whereas the compound locus composed of two cistrons will specify two distinguishable forms of the protein, isozymes, if the cistrons differ in having codons which code for amino acids of unequal charge. Such a condition has been found to exist for the alcohol dehydrogenase gene in maize. The gene which specifies this enzyme occurs singly or in duplicate. The cistrons which make up the compound locus specify different alcohol dehydrogenase isozymes. The two allelic forms found in the complex also occur singly.

Alcohol dehydrogenase in maize occurs in the developing kernel, scutellum of the mature kernel, and plumule and root of the very young seedling. In most of the work reported in this paper the scutellum was used as the enzyme source, since enzyme concentration is highest in that tissue.

TIME: _____

NONLINEAR CONDENSATION POLYMERS

Bifunctional condensation, according to the very nature of the process, necessarily leeds to products of finite molecular weight. In view of the impossibility of forcing the condensation reaction literally to completion, there will always be some few unreacted functional groups. These mark the ends of the linear molecules, which therefore are finite in length.

Nonlinear condensation polymers, on the other hand, are not restricted to growth in only two directions. It is at least conceivable that some of the molecules formed from reactants of functionality greater than two may be indefinitely large. As the polymer molecule increases in size, its functionality increases, in contrast to the linear polymers which retain only two terminal functional groups per molecule. Although some of these functional groups of the nonlinear polymer may remain unreacted, others will combine, thus continuing the structure.

Among the physical characteristics of these nonlinear condensation polymerizations, the occurrence of a sharp GEL POINT is of foremost significance. At the gel point, which occurs at a well-defined stage in the course of the polymerization, the condensate transforms suddenly from a viscous liquid to an elastic gel. Prior to the gel point, all of the polymer is soluble in suitable solvents, and it is fusible also. Beyond the gel point, it is no longer fusible to a liquid nor is it entirely soluble in solvents. Linear polymers, on the other hand, remain soluble in suitable solvents and fusible to liquids as well (unless the melting point is above the temperature of thermal decomposition), regardless of the extent of condensation.

TIME:_____

INCREASE IN DEMAND

The expression "increase in demand" will always refer to increase in the rates of consumption without an initiating reduction in price. In the tabular schedule of demand it means that the quantities are all larger for the various prices.

Confusion can be avoided by a strict observance of the difference between the amount demanded at some price and the state of demand. The latter, usually simply called demand, is the whole schedule relating prices to amounts demanded. The former, the amount demanded, is the particular amount at a specified price; it is one of the price-quantity combinations rather than the whole set of quantity-price combinations that forms the state of demand. Having noted this important distinction, we must also strictly observe the corollary distinction between changes in the amount demanded and in the state of demand. A change in the amount demanded of some good is a change in the quantity that people want to buy in response to a change in its price. This is represented by a movement along a demand schedule from one price to another. On the other hand, the concept of a change in the state of demand, called simply a change in demand, refers to a change in the amount demanded without being in response to a change in the price. With a change in demand, the amounts demanded at each possible price are different than formerly. A change in demand can be the result of a change in wealth, population, or tastes, to mention only a few of the many factors; but a change in demand is not a result of a change in price.

TIME: _____

THE WORKING OF THE POOR LAW

The working of the poor law tended to keep wages down and to perpetuate local inequalities by immobilizing some of the population. After the wars the burden of rates, in certain areas and on certain shoulders, became intolerable. There is no need to treat as representative those extreme cases, often quoted in contemporary controversy, in which the rates equalled or exceeded the annual value of the land; but it is certain that everywhere the burden hung heavy about the neck of the small holder or proprietor. A rate of 10 shillings on the acre, such as that quoted from Great Shelford, Cambridgeshire, in 1834, might well be enough to push a small man with no financial reserves over the edge of bankruptcy, in one of the bad post-war years. So the system helped to depress the "yeomanry." It worked even more disastrously on the cottager-proprietor and the "scrap-holder." The fact that he, a laborer, had to pay rates to enable other laborers to be employed at an uneconomic wage was only part of the evil. In the thoroughly pauperized areas, farmers did not care to employ such men, because no one with property was eligible for parish relief and the standard wage was so low that, without relief of some sort, it was insufficient for a married man. How numerous such cases were is not known. The principle was tragically illustrated by one put in evidence in 1824. A respectable land-owning cottager, known to be a good worker, could get no work for the reason given. His property excluded him from the "poors' books." "We must therefore wait until we are ruined," said he.

TIME:_____

INFLUENCES ON GOVERNMENT

Any attempt to trace the fundamental democratic tendencies discloses many striking contrasts between the economic, the educational and the political forces, forms and ideals. The economic evolution of the last half century tended superficially toward industrial control by the few, but fundamentally in the opposite direction. Educational development in the main was profoundly democratic, unterrified by the barriers of sex, class or race. The tendency of political institutions was left in doubt, while the evolution of political and social ideas and ideals was again distinctly democratic. The reduction of the hours of labor and universal compulsory education was democratizing influences of vast and little realized significance. One gave leisure to the adult, which meant eventually emancipation. The other developed the productive intelligence of the race, and created a basis and standard for democracy. The shorter working day and the universal educational system were powerful democratic forces, driving forward regardless of parties or bosses or governments or court decisions or of the ebbs and flows of surface opinion. If democracy as a whole could learn to think and had time to think, it was inevitable that it would in the long run draw its own democratic conclusions, whatever might happen in the short time; while ignorance and labor to the point of exhaustion on the part of the mass of the people inevitably mean rule by a few over the many regardless of what the form of government may be.

The ideals of democracy during this time were only imperfectly represented by its institutions and by their actual operation. Democratic faith was stronger than democratic works.

TIME:_____

THE COMMERCIAL BANKING SYSTEM

The manner in which commercial banks build up loans and decrease investments in prosperity, and hold down loan growth and increase investments in recessions, suggests the following conclusions concerning the role of the commercial banking system during economic fluctuations.

First, it shows that banking activities as a whole are not always effectively circumscribed by the "lock-in effect." Banks may readily decrease investments in prosperity by failing to reinvest as bonds mature or by selling off some of their holdings -- and they were apparently not afraid to do so even when some capital losses were incurred by these sales.

Second, commercial-bank loan creation in periods of recovery helps to increase demand deposits and stimulates demand, output, and employment in the economy. The data show that demand deposits have grown slowly in the last decade and a half and that most of the growth came in the year or two following each recession. These were periods of monetary ease when commercial banks had adequate reserves.

Third, when banks put their funds in investments rather than in loans during recessions, it slows down the pace of potential recovery. This is also partially responsible for the fact that demand deposits remain virtually unchanged in recessions. The money balances held by borrowers are active and stimulate the economy, but the money balances used in the purchase and sale of securities are relatively inactive.

Finally, commercial bank activity may also have an adverse effect on the economy during periods of high employment, rapidly growing gross national product, and rising prices. In these periods, banks are able to avoid monetary restrictions by reducing their investments and increasing loans.

TIME:_____

THE LOWER NUBIAN PLAIN

Alluvial or colluvial deposits floor the many shallow basins of the Lower Nubian Plain. These are mainly reddish yellow, silty coarse sands with subrounded blocky structure. Coarse sandy wash extends along the courses of the larger wadis. Eolian sand is restricted to veneers or small drifts in the lower elevation range of irregular terrain, or in the lee of desert shrubs studding some wadi beds. However, the most characteristic surface material is a lag of coarse sand and ferruginous sandstone rubble, resting on fresh or patinated bedrock. Only in rougher areas are low, free faces of bedrock exposed. These are usually patinated with ferromanganese precipitates which, in Nubia, have not yet fully discolored prehistoric rock drawings dating from the third and second millenia B.C. This preservation of patinated free faces suggests that weathering and backwearing are virtually inactive today.

Seen in perspective, the Lower Nubian Plain owes its origin to fluvial processes working during a moister phase in the Lower Pleistocene. In conjunction with our Nile Valley work, it may be attributed to wide, coalescing pediment plains cutting their way backwards from the Nile. As a result of prevailing aridity during the later Pleistocene, combined with minimal wadi gradients, subsequent activity by running water has led only to an incipient degree of dissection. Fluvial agencies cannot explain the undulating hollows, which must be attributed to wind action. Some of these closed depressions lie as much as twelve to fifteen meters below their lowest overflow thresholds, as for example the Wadi Rofa pan.

TIME:_____

8751-A

SIMPLICITY AND REFINEMENT IN WRITING

Sentiments, which are merely natural, affect not the mind with any pleasure, and seem not worthy of our attention. What an insipid comedy should we make of the chit-chat of the tea-table, copied faithfully and at full length? Nothing can please persons of taste, but nature drawn with all her graces and ornaments, or if we copy low life, the strokes must be strong and remarkable, and must convey a lively image to the mind. The absurd naivete of Sancho Panza is represented in such inimitable colours by Cervantes, that it entertains as much as the picture of the most magnanimous hero or the softest lover.

The case is the same with orators, philosophers, critics, or any author who speaks in his own person, without introducing other speakers or actors. If his language be not elegant, his observations uncommon, his sense strong and masculine, he will in vain boast his nature and simplicity. He may be correct; but he will never be agreeable. It is the unhappiness of such authors that they are never blamed or censured. The good fortune of a book, and that of a man, are not the same. The secret deceiving path of life may be the happiest lot of the one; but it is the greatest misfortune which the other can possibly fall into.

On the other hand, productions which are merely surprising, without being natural, can never give any lasting entertainment to the mind. To draw chimeras is not, properly speaking, to copy or imitate. The justness of the representation is lost, and the mind is displeased to find a picture which bears no resemblance to any original.

TIME:_____

APPENDIX B

PASSAGE DATA

Passage Means on Cloze, Comprehension,
Rate of Reading, and Preference Measures

COGNITIVE SCORES

Passage Number	Cloze	Completion Tests		Rate of Reading
		Pre-Reading	Post-Reading	
1216	551.4	473.7	576.7	184.7
1416	535.8	400.2	541.4	167.4
1821	523.8	462.2	545.2	184.3
1911	561.1	479.3	562.5	190.1
2111	518.0	416.1	547.5	176.3
2515	504.9	494.4	568.4	161.9
2725	508.9	466.6	576.6	181.8
2923	481.0	416.8	505.4	177.9
3015	500.5	423.3	518.7	185.3
3128	497.0	415.6	458.6	150.3
3213	511.5	466.4	535.9	181.9
3812	480.3	427.9	525.2	198.4
4134	469.3	403.8	470.6	174.1
4531	460.1	405.9	515.8	165.5
4814	475.6	419.5	493.3	151.7
4928	471.0	402.2	481.3	169.3
5025	446.6	417.9	477.4	161.7
5226	452.6	386.7	443.4	159.4
5636	450.8	378.2	487.9	144.2
5932	468.3	427.2	489.2	144.7
6228	439.5	368.6	428.2	150.3
6331	416.7	392.2	442.0	144.8
6441	417.2	346.1	405.4	158.7
6535	425.2	337.1	398.0	144.2
7052	379.6	326.4	342.2	139.2
7151	401.5	330.1	367.6	143.9
7453	404.6	353.3	387.9	144.3
7653	402.6	348.9	394.8	143.9
8242	373.7	347.4	371.4	128.1
8451	382.4	321.2	343.4	125.0
8552	371.6	320.4	332.0	107.5
8751	379.6	317.7	339.8	136.3

PREFERENCE RATINGS

Passage Number	Subject Matter	Textbook Reading			Reference Reading			Voluntary Reading		
		Style	Difficulty	Willingness	Style	Difficulty	Willingness	Style	Difficulty	Willingness
1216	2.78	2.58	6.06	2.74	3.06	5.82	2.88	2.32	6.06	2.48
1416	3.20	2.42	6.12	2.74	2.58	5.94	2.60	2.74	6.00	2.68
1821	4.70	4.14	4.96	4.06	4.26	5.08	4.12	3.42	5.04	3.38
1911	4.64	3.06	5.82	3.14	3.52	5.74	3.62	2.86	5.92	3.08
2111	4.48	3.72	5.12	3.42	4.12	5.04	3.82	3.54	5.38	3.22
2515	3.42	2.60	6.20	2.80	2.76	6.02	3.04	2.34	6.34	2.44
2725	4.04	3.68	6.02	3.34	3.74	5.94	3.46	3.86	5.90	3.36
2923	4.32	3.74	5.16	3.86	4.44	5.26	4.20	3.64	5.26	3.60
3015	4.28	3.86	5.50	3.80	4.26	5.48	4.08	3.68	5.62	3.58
3128	3.80	3.82	4.62	3.50	4.42	4.66	4.32	3.12	4.74	2.98
3213	4.26	4.04	5.34	3.78	4.56	5.26	4.12	3.78	5.58	3.50
3812	4.00	3.82	5.60	3.40	3.92	5.18	3.66	3.12	5.62	3.08
4134	4.36	4.56	4.40	3.82	4.60	4.42	4.10	3.66	4.50	3.22
4531	4.42	4.40	4.60	3.68	4.58	4.54	3.90	3.70	4.64	3.52
4814	3.78	3.46	4.82	3.14	4.16	4.72	3.52	2.92	4.84	2.68
4928	4.08	4.08	4.32	3.70	4.76	4.58	4.18	3.76	4.52	3.32
5025	4.82	4.18	5.26	3.68	4.32	5.04	4.04	3.70	5.38	3.50
5226	4.36	4.00	4.36	3.32	4.22	4.44	4.10	3.38	4.20	2.96
5636	4.26	4.48	4.50	3.82	4.60	4.76	4.28	4.06	4.50	3.58
5932	4.76	4.40	4.66	4.10	4.60	4.72	4.16	4.00	4.96	3.68
6228	4.02	3.90	3.66	3.38	4.42	3.96	3.94	2.94	3.78	2.70
6331	4.18	3.96	3.90	3.84	4.64	4.02	4.52	3.36	3.70	3.40
6441	4.26	4.18	3.76	3.44	4.48	4.12	4.06	3.56	3.92	3.36
6535	4.12	4.18	3.70	3.52	4.42	4.02	4.02	3.58	3.78	3.30
7052	3.54	3.36	2.56	3.18	4.14	2.88	3.68	3.16	2.50	3.04
7151	2.68	3.20	2.36	2.64	3.46	2.76	3.00	2.72	2.42	2.42
7453	3.38	3.42	3.08	2.92	3.56	3.26	3.42	2.62	3.38	2.60
7653	3.70	3.38	3.34	3.30	4.06	3.60	3.82	2.64	3.34	2.64
8242	3.76	3.66	3.14	3.24	4.22	3.46	3.98	2.94	2.94	2.82
8451	3.54	3.08	3.18	2.52	3.58	3.30	3.18	2.20	2.92	2.28
8552	3.46	3.42	2.58	3.04	3.84	2.94	3.58	2.76	2.76	2.70
8751	3.38	3.14	3.18	2.78	3.74	3.38	3.24	2.44	3.16	2.50

APPENDIX C

INTEREST SCALES

RATING SCALES

1. How well do you like to learn about the subject this book talks about?

Subject:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

When used as your school textbook this year.

2. Interest:

1	2	3	4	5	6	7
Very dull	Dull	A little dull	Neither dull nor interesting	A little interesting	Interesting	Very interes

3. Level:

1	2	3	4	5	6	7
Much too hard	Too hard	A little too hard	About right	A little too easy	Too easy	Much too easy

4. Preference:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

When used as one of your reference books this year.

5. Interest:

1	2	3	4	5	6	7
Very dull	Dull	A little dull	Neither dull nor interesting	A little interesting	Interesting	Very interes

GO TO NEXT PAGE IMMEDIATELY

RATING SCALES

6. Level:

1	2	3	4	5	6	7
Much too hard	Too hard	A little too hard	About right	A little too easy	Too easy	Much too easy

7. Preference:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

When used as one of your library books this year.

8. Interest:

1	2	3	4	5	6	7
Very dull	Dull	A little dull	Neither dull nor interesting	A little interesting	Interesting	Very interest

9. Level:

1	2	3	4	5	6	7
Much too hard	Too hard	A little too hard	About right	A little too easy	Too easy	Much too easy

10. Preference:

1	2	3	4	5	6	7
Dislike very much	Dislike	Dislike a little	Neither like nor dislike	Like a little	Like	Like very much

STOP. DO NOT TURN THE PAGE.

APPENDIX D

INSTRUCTIONS

2/20/70

INSTRUCTIONS FOR ADMINISTERING THE TESTS

1. *Introduce yourself to the teacher:*

   a. *Tell him who you are and why you are there.*

   b. *Thank them for letting us give the tests.*

   c. *Write on board:*

   *Name*_____
   *School*_____
   *Grade*_____
   *Room*_____

2. *Settle the class*

3. *Announce:*

We are conducting a study and we need your help. We want to know what kinds of books you are able to learn the most from and what kinds of books you like to study best.

Does everyone have a pencil?

4. *Pass out the booklets and say:*

Now, I am going to pass out some booklets containing some tests which you are asked to take. As soon as you get your booklet, please fill out the blanks on the front. DO NOT OPEN IT UNTIL I TELL YOU TO DO SO.

Your name goes in the first blank, your school's name goes in the second, your grade in the third blank, and this room's number in the last blank.

Do NOT open your booklet until I tell you to.

(Wait for them to finish filling in the blanks.)

There are four different kinds of tests for you to do in your booklet. Everyone will take the same kinds of tests but the student next to you will probably be reading about something different than you will be. Your tests may be quite difficult, easy, or just right for you. It is important for you to do your best work no matter what the difficulty of the tests.

Now turn to the first page. D-1

5. *Instructions for the Cloze Test:*

This test was made by taking every fifth word out of a story. A blank was left where a word was taken out. You are to write in each blank the word you think was left out.

Most of the blanks can be answered with ordinary words but a few may be
numbers like . . . . . 3,427 or $12 or 1954.
contractions like. . . . . can't or weren't
abbreviations like . . . . . Mrs. or U.S.A.
parts of hyphenated words like . . . . . self in the word self-mad
But most will be just ordinary words.

You are to write only <u>one</u> word in each blank.
Do NOT be afraid to guess. Wrong spelling will <u>not</u> count against you if we can tell what word you meant. I will help you spell, if you raise your hand.

Try to write an answer for every blank. But don't waste too much time on a blank. Some blanks are very hard. You should skip them and then try them again when you have finished.

Please work rapidly but carefully. And remember, do NOT write more than one word in each blank. And please write neatly.

When you finish, just leave your booklet open to this page and put your pencil down. Remember, you may <u>not</u> go ahead to the next page. Any questions? You may begin.

TIME: Elementary Schools - - - - - - - - - - - 16 minutes
Junior and Senior High Schools - - - - 11 minutes

6. *Supervise the testing:*
   a. *Circulate in the room*
   b. *Help children spell words, but ONLY if they first tell you the word they want to spell. Then write it down for them.*
   c. *Do NOT pronounce words for students or tell them what the words mean.*
   d. *Encourage the students to work rapidly once or twice early in the test by*
      (1) *At least twice during the testing advise the students to skip the hard ones and come back to them later. Do this in a voice just loud enough to be heard by all and in a way that does not interrupt their work.*
      (2) *Announce quietly that the students should work carefully but quickly.*
      (3) *About halfway through the testing period announce quietly that they should be about half done.*
      (4) *About two minutes from the end of the testing period announce that they will have a minute or two to finish.*
   e. *Keep all announcements made during testing in a soft tone just loud enough to be heard by all students.*
   f. *Because we will be working with children of vastly differing ability level you will have to "play it by ear" in stopping the test. Here are a few rules which must be observed:*
      (1) *Watch to see when they have answered the last few questions.*
      (2) *Watch to see when they have stopped writing responses.*
      (3) *Watch to see when all have finished productive work; when just a few students are left who do not seem to be working productively, advise them that they have a couple of minutes to complete the test. Wait a minute or so and then stop the test.*
      (4) *Do not exceed the time limits. Two minutes prior to the end of the time limit advise the students that they have a couple of minutes to complete the test.*
      (5) *Stop the test by saying, "Please stop and turn to the next page."*

7. Instructions for the completion guessing test.

Now let's turn to the next page. The purpose of this test is to help us find out how much you already know about a story even before you have studied it. Read each question and try to answer it as well as you can even though you have not read the story the test was made for. The correct answer may be only a single word; it may be a few words; or it may be a sentence or two. Sometimes the same answer may be given to more than one question.

Don't be afraid to guess at the answers. Try to write your very best guess for each question. If you find that a question is too hard on your first try, skip it and return to it when you have finished the other questions. But try to write a good answer for every question that you can.

There are 20 questions in this test and they are printed on two pages. When you finish the first page, go directly to the next page. Do NOT wait for me to tell you to turn the page. When you have finished the second page, you may go back and check your work on this test. But do not go on to the next test or back to the previous test. Any questions? Begin.

TIME: Elementary Schools - - - - - - - - - - - - - - - 13 minutes
Junior and Senior High Schools - - - - - - - - 11 minutes

8. Supervise the completion guessing test:

a. Follow essentially the same procedures as you followed on the cloze test.
b. About halfway through the period announce that they should now be working on the SECOND PAGE of the questions.
c. Advise them not to waste too much time on the hard questions.
d. Use the same procedure as you used on the cloze tests for stopping the test.

9. Instructions for Administering the Rating Scales:

Now turn to the next test. Notice that the next two pages are blank. Turn past them to the page that has the story on it. On the left is a story taken from a textbook. On the right-hand side you see the first page of a set of rating scales. The story was selected because it could give you a very good idea of what the rest of the book was like. First read it and then you will be asked to rate it to show how suitable you think this book would be for you personally to use for three different purposes.

First, you will be asked to rate how well you personally like to learn about the subject talked about in the passage. Then you will be asked to say how well you would like to use this textbook this year as the regular textbook for one of your classes, as a reference book just for looking up special things, and as a library book you might read just for pleasure in your spare time. You will be asked to say whether it is interestingly written and at the right level of difficulty for you personally for each of these purposes.

First you should read the passage and then I will explain each question to you BEFORE you mark your response. Do not try to answer a question until I have explained it to you. Any questions? Read the passage.

TIME: Elementary Schools - - - - - - - - - - - - - - - - - 3 minutes, 30 seconds
Junior and Senior High Schools - - - - - - - - - - 3 minutes

10. Administration of the Rating Scales:

a. Circulate to see that no one starts to answer the questions before you tell them to.
b. Follow the usual rule about not pronouncing words for students.
c. Begin with the scales before the time limit is out if it appears that nearly all have finished reading the passage.

11. The Rating Scale Questions:

Now look at the scales on the right hand page. I will read and explain each question and tell you how to mark it. Stay with me as we go through the scales.

Your answers to these questions should show us exactly how you personally would feel about using this textbook--not how your older brother or younger sister might feel about the book. Your answers should also tell us how you would feel about using the book this year--not how you might have felt a year or two ago or a year or two in the future.

Now look at question number 1. It asks how well you like to learn about the subject matter talked about in this book. On this question, do not try to rate the book, itself. Rather, we want to know just how much you like to learn about this subject--the kind of things this book talks about, regardless of whether you learn it by talking to others, watching T.V., reading a book, or listening to a teacher.

You will show how much you like to learn about this subject by circling one of the numbers on the scale below the question. Notice that the words under each number tell what the number means.

(Hold up your book and demonstrate)

If you DISLIKE IT VERY MUCH, you should draw a circle around the number 1.
Circling the 2 means you DISLIKE IT.
3 means you DISLIKE IT A LITTLE.
4 means you NEITHER LIKE NOR DISLIKE IT.
5 means you LIKE IT A LITTLE.
6 means you LIKE IT.
7 means you LIKE IT VERY MUCH.

Now circle the number which shows how well you like to learn about the subject, the kinds of things, this book talks about.

(Pause a few seconds and glance around to assure that everyone knows what he is supposed to do)

Next, you are asked to answer three questions about how you would feel about having to study this book, first, as the regular textbook in one of your classes, then as a reference book for looking up special things, and finally as a library book that you read just for pleasure in your spare time.

The first set of three questions ask how you would like to study this book as the regular text book used in one of your classes.

Question number 2 asks, How interestingly do you think this book was written-not what the author talked about but how he talked about it--if you had to use it as the regular textbook in one of your classes.

On this scale number 2 notice that

1 means VERY DULL
2 means DULL
3 means A LITTLE DULL
4 means NEITHER DULL NOR INTERESTING
5 means A LITTLE INTERESTING
6 means INTERESTING
7 means VERY INTERESTING

Now circle the number which shows how interestingly you think this book was written--not what the author talked about but how he talked about it. You m take a few seconds to glance back at the passage, if you need to. Remember that all of your answers are supposed to show just how you personally would feel about using this book this year.

(Allow about 20 seconds on this)

Circle your response for question 2.

(Pause very briefly)

Next look at scale number three. Your response on this scale will show how hard or easy you think this book is for you, if you had to use it as the regular textbook in one of your classes.

On scale number 3 notice that

1 means MUCH TOO HARD
2 means TOO HARD
3 means A LITTLE TOO HARD
4 means ABOUT RIGHT
5 means A LITTLE TOO EASY
6 means TOO EASY
7 means MUCH TOO EASY

Circle the number which shows how hard or easy you think this book is for you, if you had to use it as the regular textbook in one of your classes. You may glance back at the passage, if you need to.

(Pause about 10 seconds)

Circle your response.

Question number 4 asks how well you would like (or how willing you would be) to read this book, if you had to study it as the regular textbook in one of your classes. This scale is like the one used for question number one. You may glance back at the story if you need to.

(Pause about 10 seconds)

Mark your response.

The next three questions are the same as questions 2, 3, and 4 except that these questions ask you to rate the book on how well you would like to use the book just as a reference book for looking up special things.

Question 5 asks you to circle the number which shows how interestingly written you think this book is (not what the author talked about but how he talked about it) if you had to use it as a reference book for looking up special things.

Mark your answer.

Question number 6 asks you to circle the number which shows how hard or easy you think this book is for you if you just had to use it as a reference book for looking up special things.

Mark your answer.

Question number 7 asks you to circle the number which shows how well you would like (or how willing you would be) to read this book if you just had to use it as a reference book for looking up special things.

Mark your answer.

The next three questions are just like the last three except that these new questions ask you to rate the book on how well you would like to use it as a library book which you read just for pleasure in your spare time.

Question 8 asks you to circle the number which shows how interestingly you think this book was written (not what the author talked about but how he talked about it) if you just had to use it as a library book you would read just for pleasure in your spare time.

Mark your response.

Question 9 asks you how hard or easy you think this book is for you, if you just had to use it as a library book you would read just for pleasure in your spare time.

Mark your answer.

Question 10 asks how well you would like (or how willing you would be) to read this book, if you just had to use it as a library book you would read just for pleasure in your spare time.

Mark your answer.

12. *Instructions for Administering the Postreading Completion Test:*

Do not turn the page until I tell you to. Next, you are going to read a story and take a test that is about that story. This test is like the second one you took except that this one asks questions about the story you are going to read. Your answers on this test may be a single word, a few words, or even a sentence or two. Sometimes the same answer may be given to more than one question.

You will also mark down the amount of time it took you to read the story. Here on the chalkboard *(point)* I have written down some times. I will watch the clock while you are reading and erase these numbers one at a time. When you finish reading the story, look up here and write down whatever number is at the top at that time. Write this number in the empty space below the story. Then turn immediately to the questions. But this is not a speed test. Your main job is to learn what the story talks about because you may not look back at it later.

When you turn to the questions, don't be afraid to guess at the answers. Try to write your very best guess for each question. If you find that a question is too hard on your first try, skip it and return to it when you have finished the others. But try to write a good answer for every question that you can.

There will be 20 questions in the test, and they are printed on two pages. When you have finished the first page, GO DIRECTLY TO THE NEXT PAGE. Do NOT wait for me to tell you to turn the page. When you have finished the second page, you may go back and check your work on this test. But do NOT go back to the story or to the previous tests.

Remember:

1. Read the story carefully.
2. As soon as you finish reading it, look up here immediately to see how long it took you to read it.
3. Write the number that is at the top of the list below the story.
4. Then go immediately to the test.
5. DO BOTH PAGES OF QUESTIONS.

Any questions? Turn the page and begin reading the story.

Version for measuring rate

TIME: *There are no time limits on this test but you can expect the students to have all finished in roughly these times:*

Elementary School..... 18 minutes

Junior and Senior High School..... 13 minutes

13. *Administration of the Postreading Completion Test:*

a. *Write the following numbers on the chalkboard in vertical columns:*

1:00, 1:10, 1:20, 1:30, 1:40, etc. to 6:00

b. *While they are reading, mark your time from when you told them to turn the page and begin to read and then erase the number at the top of the list as that time is reached. That is, when one 1:00 minutes have elaps erase that number.*

c. *Watch the students as they finish to be sure that they are looking up t see what their time was and that they are writing it down.*

d. *Remind them quietly three or four times to note down their time as they finish reading.*

When they have all finished reading:

a. *Circulate in the room to see that the instructions are followed.*

(1) *See that they do BOTH pages of questions.*

(2) *They should turn directly to the tests without waiting for you to tell them.*

(3) *They should not turn back to the passage after they have turned t the questions.*

b. *Encourage them to work carefully but rapidly in the same manner as bef*

14. *Fi h the testing:*

a. *Tell the students to close their books.*

b. *Have them check to see that they filled out the front page of the b-ok completely and correctly.*

c. *Ask for any booklets which seem to be falling apart.*

d. *Pick up the booklets.*

e. *Band and label the stack of booklets with school name, grade, and room*

f. *Thank the kids (and the teacher).*

g. *If you have time, answer any questions about the study.*

12. Instructions for Administering the Postreading Completion Test:

Now turn to the next page where you will see another story. First, you will read this story and then you will take another test. This test is like the second one you took except that this one asks questions about the story you are about to read. Your answers on this test may be a single word, a few words, or even a sentence or two. Sometimes the same answer may be given to more than one question.

First, read this story carefully and then turn immediately to the test and begin answering the questions. Answer the questions just from what was contained in the story. Once you have turned to the test you must NOT turn back. So study the story carefully before you turn to the test.

Don't be afraid to guess at the answers. Try to write your very best guess for each question. If you find that a question is too hard on your first try, skip it and return to it when you have finished the other questions. But try to write a good answer for every question that you can.

There are 20 questions in this test and they are printed on two pages. When you finish the first page, go directly to the next page. Do NOT wait for me to tell you to turn the page. When you have finished the second page, you may go back and check your work on this test. But do not go back to the story or to the previous test.

Any questions? Begin.

TIME: Elementary School - - - - - - - - - - - - - - - - - 18 minutes
Junior and Senior High School - - - - - - - - - - 13 minutes

13. Administration of the Postreading Completion Test:

a. Circulate in the room to see that the instructions are followed.
(1) See that they do BOTH pages of questions
(2) They should turn directly to the tests without waiting for you to tell them.
(3) They should not turn back to the passage after they have turned to the questions.

b. Encourage them to work carefully but rapidly in the same manner as before.

14. Finish the testing:

a. Tell the students to close their books.

b. Have them check to see that they filled out the front page of the booklet completely and correctly.

c. Ask for any booklets which seem to be falling apart.

d. Pick up the booklets.

e. Band and label the stack of booklets with school name, grade, and room numbe:

f. Thank the kids and the teacher.

g. If you and the teacher have time, answer any questions about the study.

SUPPLEMENT TO THE TEST MANUAL

This supplement contains two major parts. The first is a general explanation of the purpose of the study. It should be read to the students at the beginning of the testing. The second is an abbreviated version of the instructions in the test manual. This abbreviated version should be used in place of the version in the manual after the second or third testing session when the students have become well acquainted with the testing procedures.

PURPOSE OF THIS STUDY

You have been selected to participate in an important study. The purpose of this study is to find out what kinds of textbooks help students learn the most. Some students have difficulty with their schoolwork because their textbooks are too hard for them. And we could help those students a great deal if we just provided them with books that were at the proper level of difficulty. Schools would like to do this, if they could. But they have not been able to do it very well because scientists have not yet found out what level of difficulty is actually best for a student. That is what this study is designed to find out.

ABBREVIATED INSTRUCTIONS

4. *Pass out the booklets and say:*

Check to be sure you get the booklet with your name on it.

5. *Instructions for the Cloze Test:*

Now turn to the first test. Write in each blank the word you think was taken out. Remember, write just one word in each blank and don't be afraid to guess. The correct answer is usually a word but some may be numbers, contractions, abbreviations, or words with hyphens in them. Try every blank and don't be afraid to guess. Any questions? Begin.

7. *Instructions for the Completion Guessing Test:*

Turn to the next page. This is the test where you are just supposed to guess at the answers. Make sure you give your very best guess and try to write a guess for every question. Your answers to these questions may be as long as necessary. There are two pages of questions. Don't stop till you have done both pages. Any questions? Begin.

9. *Instructions for Administering the Rating Scales:*

Now turn to the next page. This is the story you will be asked to read. Read it carefully and then wait for me to read you the questions before you try to answer them.

Now let's answer the first question, question number 1.

Circle the number that shows how well you like to learn about the subject matter talked about in this book, regardless of whether you learn it by talking to others, watching T.V., reading a book, or listening to a teacher.

The next three questions ask you to rate this book when it is used as the regular textbook in one of your classes. For question Number 2, circle the number that shows how interestingly you think this book was written--not what it talks about but how it talks about it--if you had to use it as a regular textbook in one of your classes.

Number 3: Circle the number that shows how hard or easy you think this book is for you, if you had to use it as the regular textbook in one of your classes.

Number 4: Circle the number that shows how well you would like (or how willing you would be) to read this book, if you had to study it as the regular textbook in one of your classes.

The next three questions ask you to rate this book when it is used just as a reference book for looking up special things.

Number 5: Circle the number that shows how interestingly written you think this book is (not what the author talks about but how he talked about it) if you had to use it just as a reference book for looking up special things.

Number 6: Circle the number that shows how hard or easy you think this book is for you if you had to use it just as a reference book for looking up special things.

Number 7: Circle the number that shows how well you would like (or how willing you would be) to read this book if you had to use it just as a reference book for looking up special things.

The next three questions ask you to rate this book when it is read as a library book which you read just for pleasure in your spare time.

Number 8: Circle the number that shows how interestingly written you think this book is (not what the author talked about but how he talked about it) if you just had to use it as a library book you would read just for pleasure in your spare time.

Number 9: Circle the number that shows how hard or easy you think this book is for you, if you just had to use it as a library book you would read just for pleasure in your spare time.

Number 10: Circle the number that shows how well you would like (or how willing you would be) to read this book, if you just had to use it as a library book you would read just for pleasure in your spare time.

12. *Instructions for Administering the Postreading Completion Test:*

Do not turn the page until I tell you to. On the next test you will

First, read the story.
Second, when you finish look up here and note the time it took you to read it.
Third, write that number in the empty space below the story.
Fourth, turn to the questions and answer both pages of questions.

Remember, this is not a speed test so study the passage as well as you need to. But don't forget to mark down how much time it took you to read it. And answer the questions on BOTH pages, but do not look back at the story.