DOCUMENT RESUME

ED 082 138 CS 000 728

AUTHOR MacGinitie, Walter H., Ed.
TITLE Assessment Problems in Reading.
INSTITUTION International Reading Association, Newark, Del.
PUB DATE 73
NOTE 107p.
AVAILABLE FROM International Reading Association, 6 Tyre Avenue, Newark, Del. 19711 (Order No. 462, $3.00 non-member, $2.00 member)

EDRS PRICE MF-$0.65 HC-$6.58
DESCRIPTORS Classroom Environment; *Criterion Referenced Tests; Reading; *Reading Diagnosis; Reading Instruction; Reading Materials; Reading Processes; Reading Skills; *Reading Tests; *Standardized Tests; *Test Interpretation; Test Results

ABSTRACT

The papers in this volume deal with a range of assessment problems in reading. The first paper, by Karlin, introduces the general problem of using assessment procedures to guide teaching. The next six papers deal with various aspects of this general problem. Otto discusses the distinction between norm-referenced, standardized achievement tests and criterion-referenced measures. Johnson shows how the teacher can prepare his own criterion-referenced evaluation procedures to fit specific objectives in word attack skills. Berg's paper documents the difficulty in evaluating specific components of reading ability. MacGinitie points out that the nature of what is tested in reading changes from the lower to the higher grades. Carver critically analyzes the relationship between reasoning and reading. Thorndike discusses some of the problems of test interpretation. The next two papers deal with the instructional setting and the instructional materials. Brittain provides a checklist of points to consider when evaluating classroom organization. A paper by Botel, Dawkins, and Granowsky offers a way of analyzing the structures of sentences to estimate their complexity. The last two papers, Mork's and Jason and Dubnow's, consider the relationship between the reading ability of the child, the material he reads, and his assessment of his reading ability. (WR)

assessment problems in reading

Walter H. MacGinitie, *Editor*
Teachers College
Columbia University

INTERNATIONAL READING ASSOCIATION
Six Tyre Avenue • Newark, Delaware 19711

INTERNATIONAL READING ASSOCIATION

OFFICERS

1973-1974

President Millard H. Black, Los Angeles Unified School District, Los Angeles, California

President-Elect Constance M. McCullough, California State University, San Francisco, California

Past President William K. Durr, Michigan State University, East Lansing, Michigan

---

DIRECTORS

Term expiring Spring 1974

William Eller, State University of New York, Buffalo, New York
William J. Iverson, Stanford University, Stanford, California
Eunice Shaed Newton, Howard University, Washington, D. C.

Term expiring Spring 1975

Harold L. Herber, Syracuse University, Syracuse, New York
Helen K. Smith, University of Miami, Coral Gables, Florida
Grace S. Walby, Child Guidance Clinic of Greater Winnipeg, Winnipeg, Manitoba

Term expiring Spring 1976

Ira E. Aaron, University of Georgia, Athens, Georgia
Lynette Saine Gaines, University of South Alabama, Mobile, Alabama
Tracy F. Tyler, Jr., Robbinsdale Area Schools, Robbinsdale, Minnesota

---

Executive Secretary-Treasurer Ralph C. Staiger

Assistant Executive Secretaries Ronald W. Mitchell
James M. Sawyer

Business Manager Ronald A. Allen

Director of Research Stanley F. Wanat

Journals Editor Lloyd W. Kline

Publications Coordinator Faye R. Branca

Library of Congress Catalog Card Number 73-84793

CONTENTS

The International Reading Association attempts, through its publications, to provide a forum for a wide spectrum of opinion on reading. This policy permits divergent viewpoints without assuming the endorsement of the Association.

FOREWORD

The title of this publication, *Assessment Problems in Reading,* reflects many of the present concerns and the future hopes of reading instruction. Failures to accurately assess both pupil needs and instructional objectives are among the causes of educational ineffectiveness. Methods by which teacher skill in these and related areas may be increased are matters of administrative and legislative concern in many parts of this country and the world.

The need for assessment is present in every important area of the instructional program. What are the strengths and weaknesses of the pupil, the teacher, the school, and the instructional support program? How may the effectiveness of classroom organization be increased? How accurately has the instructional level of the pupils been determined? How appropriately do the materials of instruction reflect the abilities and the needs of the pupils? How do the pupils perceive themselves, the teacher, and the educational process?

Evaluation is in a period of crisis and change. Parents, teachers and administrators of public schools, college and university personnel, and the critics of education in general are questioning the validity of time-honored evaluation procedures. What is the impact of an evaluation program on the pupils it is designed to serve? Do test scores illuminate and guide, or do they obfuscate and confuse? It is difficult to conceive of effective teaching without procedures for determining the skills the pupil possesses, the ways in which his needs are similar and dissimilar from those of his

peers, the degree to which his educational progress parallels that of other pupils of his age and grade.

Similarly, an effective teacher is perceived as selecting from among the wide range of instructional media in terms of the established needs of each pupil in his class and choosing those materials to meet the needs and to reinforce the strengths of each of these young people.

The Association is indebted to Walter MacGinitie and the authors of ***Assessment Problems in Reading*** for the contribution they have made in the preparation of this publication in order that the teaching of reading may be improved through more effective assessment of the many aspects of the instructional process.

Millard H. Black, ***President***
International Reading Association
1973-1974

Walter H. MacGinitie
Teachers College
Columbia University

AN INTRODUCTION TO SOME MEASUREMENT PROBLEMS IN READING

The papers in this volume deal with a wide range of assessment problems in reading. The first paper by Karlin introduces the general problem of using assessment procedures to guide teaching: setting appropriate objectives for the child and selecting assessment procedures that will contribute to an understanding of the child's current capabilities and of appropriate new goals. The next seven papers deal with various aspects of this general problem. Otto discusses the distinction between norm-referenced, standardized achievement tests and criterion-referenced measures, and makes clear the distinctive usefulness of the latter. Johnson shows how the teacher can prepare his own criterion-referenced evaluation procedures to fit specific objectives in word attack skills. Beyond the decoding stage, it has been more difficult to develop ways to evaluate specific components of reading ability or even to identify these components clearly. Berg's paper clearly documents this difficulty. MacGinitie points out that the nature of what we teach in reading, and therefore the nature of what is tested, changes from the lower to the higher grades. Carver takes issue with the desirability of such a change and describes what he believes to be the undesirable consequences of an emphasis on teaching and measuring the reasoning aspects of reading in the later grades.

All of the foregoing papers are concerned to some degree with differential measurement, that is, discovering whether different aspects of reading ability can be distinguished and measured and, if they can, how those differential measurements can be used in

guiding instruction. The paper by Thorndike provides a remarkably clear discussion of the basic measurement problems inherent in trying to learn if a particular child is better at one task than another. His clear description of the statistical relations involved, and the simple tables that he provides for guiding diagnostic judgments, should be invaluable both in making diagnoses and in evaluating the usefulness of diagnostic instruments.

The next two articles turn from measuring student achievement to look at the instructional setting and the instructional materials. Brittain's discussion and checklist of points to consider when evaluating classroom organization will be useful not only to the school that is planning such an evaluation but also to the individual teacher who simply wants to think through what he would like to accomplish through organizing his classroom for reading instruction. That we can scale the reader's ability suggests that we can scale the difficulty of the material he reads, and indeed there are many procedures for assessing readability. Most readability formulas use sentence length as an estimate of sentence complexity. The paper by Botel, Dawkins, and Granowsky offers a relatively simple way of analyzing the actual structures of sentences to achieve estimates of their complexity.

The last two papers in this volume consider the relationship between the reading ability of the child, the material he reads, and his own assessment of his reading ability. Mork inquires whether children can select materials that are appropriate for their level of reading achievement, and Jason and Dubnow report a study of the relation between reading achievement and children's assessments of their own reading ability.

One of the most persistent of the many issues raised by the papers in this volume involves the reliability of difference scores. Several of the papers stress the importance of diagnostic testing or diagnostic judgments. It is important to good teaching to realize how fallible such differential test results or judgments ordinarily are, so that instructional decisions can be kept appropriately tentative. Most reading skills—especially the more advanced comprehension skills—are highly correlated with one another, and only when the subskill scores are quite different from each other can diagnostic judgments of practical usefulness be made.

Since a teacher's judgments are likely to be at least as unreliable and as highly intercorrelated as test subscores are, the sober-

ing message of Thorndike's tables applies in full measure to teachers' judgments as well. The teacher should be at least as tentative about diagnostic judgments formed from his own observations as those formed from test scores and be ready to change both appraisal and treatment as new evidence warrants. The references in several articles to the value of systematic observation, anecdotal records, and teacher-made diagnostic instruments suggest the need for more training in these skills in teacher education.

Some of the papers in this volume make clear the advantages that criterion-referenced tests have over norm-referenced tests for certain purposes, particularly for guiding teaching. It should be understood, however, that criterion-referenced measures are often used for making diagnostic judgments and are subject to related limitations for that purpose. Giving a score that refers to some criterion rather than to a norm group does not absolve the test maker from showing that separate component scores index meaningful skill levels or separately measurable skills. Whether the test is criterion-referenced or norm-referenced, the teacher must recognize that the label of the test is not necessarily a clear guide to what the test measures. The problem of subtests that have different labels but that do not actually measure different skills is ably described in the article by Berg. Unless criterion-referenced tests clearly demonstrate that they are relevant to different criteria, they are likely to perpetuate the same problem. In evaluating the distinctive contributions of criterion-referenced and norm-referenced tests, it is well to remember that both types are usually standardized in the sense that they are given with standard directions and under standard conditions. For example, the first four of Otto's Limitations of Standardized Tests can apply to criterion-referenced as well as to norm-referenced tests. Finally, it is just as important in using criterion-referenced tests as in using norm-referenced tests to be sure that the test that is used is appropriate to the objectives that are guiding the instruction.

Two of the papers make reference to the fact that grade equivalents obtained from standardized, norm-referenced reading tests do not provide a very accurate index of the child's instructional level. That a discrepancy exists is quite true, but the reasons for it seem not be be generally understood. First of all, the child's instructional level, as determined by an informal reading inventory, is usually based on some graded series of reading texts or on

standardized test passages. However, the materials for a particular grade level produced by one publisher may be considerably harder or easier than those produced by another publisher, and standardized test passages used for determining instructional level have no inherent priority over other standardized test passages. For these reasons the instructional level as determined by one informal reading inventory may not agree with a second using different materials.

Secondly, the traditional criteria for instructional level (96 percent correct pronunciation and 60 percent comprehension) are quite arbitrary and not always comparable to each other. Furthermore, the comprehension score depends on the questions asked, and it is clear that questions of varying difficulty may be asked about the same passage. The arbitrariness of the comprehension criterion for instructional level is particularly evident. Why does answering 60 percent of the questions about a passage mean that passage is appropriate in difficulty for the child to study? And, indeed, is the same criterion appropriate at all grade levels?

Finally, instructional level and the grade score from a test are based on opposite regression lines.* For this reason, adding or subtracting a constant will not, as implied by some recent investigators, serve to convert different grade scores to corresponding instructional levels. The grade score will often give a fair indication of instructional level, but the grade score is not defined in such a way as to give the best estimate of the level of reading material most appropriate for the child. By convention, the grade scores from a reading test are based on the average raw scores obtained by children at each of several different grade levels. The line through the mean raw score points at different grade levels is essentially the regression line for the regression of raw score on grade level (*1*). If, on the other hand, one were trying to predict the grade level of a child who has received a particular raw score,

*The situation is actually more complex than this description indicates, and revising the conventional definition of the grade equivalent would not provide a thoroughly satisfactory answer. A practical solution to the problem is complicated by the fact that raw scores should actually be plotted against appropriate level of instructional material, rather than the actual grade level of the student, and separate regression solutions should be obtained for students of each actual grade level. The description does not specify whether or not the regression is linear, but that is not a relevant consideration for the point being made. The line through the mean raw score points at different grade levels is usually curvilinear, with smaller and smaller increments from year to year as grade level increases.

one would be interested in the average grade level of all the children in the normative sample who received that particular raw score. If raw scores were assigned grade level equivalents on this latter basis, the assignment would be based on the opposite regression line—the regression of grade levels on raw scores.

Three of the articles in this collection merit special comment—the paper by Carver because it represents a novel approach to the measurement of reading, and the papers by Mork and by Jason and Dubnow because they represent important beginnings in promising areas of research.

As background for understanding Carver's article, it would be very helpful for the reader first to read E. L. Thorndike's original article, "Reading as Reasoning," reprinted in a recent *Reading Research Quarterly* (2) and Tuinman's perceptive commentary on it (3). Carver's intriguing article was solicited by the Editor, recognizing that the article would be controversial, but believing that it was appropriate to give wider currency to Carver's thought-provoking views on the nature of reading instruction and reading measurement. Carver attacks the generally accepted concept of the nature of reading as exemplified by some of the work of E. L. Thorndike and R. L. Thorndike. Since there is no reply to Carver contained in this volume, the Editor wishes to defend the elder Thorndike on one specific point and to suggest very briefly the nature of some of the questions that a general defense might raise. Carver maintains that E. L. Thorndike was interested primarily in studying the decoding of words and the combining of word meanings into an understanding of the sentence, but objects that Thorndike's actual work did not concentrate on those levels. In view of the subtitle ("A Study of Mistakes in Paragraph Reading") of Thorndike's major article, it seems inappropriate to take Thorndike to task for allowing his interest to range beyond the sentence. The types of questions that a defense of the Thorndikian view might raise are exemplified by the following: Is the teacher happy with a definition of the reading process that specifically excludes meanings that go beyond those contained in a single sentence? Is the distinction between understanding a sentence and understanding a paragraph a valid one? If Carver's view is correct, how can one explain the high correlations between knowledge of individual words on a vocabulary test and scores on a comprehension test? Is not reading, as we teach it, intended to be a useful skill so that a

person who has read something should be able to do something that he could not do before (for example, answer a question about what he has read)?

The articles by Mork and by Jason and Dubnow are closely related, as they both deal with the question of what the child understands about his own reading ability. Mork shows that many children in the third and fifth grades are able to select reading materials that are appropriate for their level of reading ability. Many other children, however, select materials that are considerably too easy or too difficult, according to a readability analysis of the selected material. The basic question is obviously a good one, and Mork plans additional studies to investigate how well children can actually read the particular materials that they select. The possibilities for studying the effects of interest, motivation, and specific subject matter are clearly important.

The paper by Jason and Dubnow is also concerned with children's evaluations of their own reading ability. The authors' underlying concern is how the child's perception of himself as a reader influences his growth in reading ability. This initial research clearly shows a relation between the child's perception of his reading ability and his tested reading achievement. One way of interpreting these results is that a child has a pretty fair idea of how well he reads, an interpretation that conforms with Mork's findings. An alternative possibility, and one that motivates Jason and Dubnow's work, is that the child's perception of himself as a reader has actually influenced his development as a reader. The present study does not allow one to choose between these two interpretations, but the problem is an important one and could be studied by causing changes either in the perception or the ability.

The question of how reading self-concept *can* be changed is in itself an interesting basis for research. At the end of their report, the authors draw attention to another very interesting possibility: that the child's own perception of his reading strengths and weaknesses may be valuable information for planning individualized or remedial teaching. Does the child have any diagnostic awareness of his reading capabilities or only a global evaluation? If he does have some sort of diagnostic awareness, what is his taxonomy of the reading task? Could children's unstructured descriptions of their own specific strengths and weaknesses in reading contribute to our understanding of the process of learning to read?

These questions illustrate how a research report can provide helpful answers to one problem and also raise new problems leading to new explorations that increase our understanding of reading. The Editor hopes that each of the articles in this collection will serve these two purposes of clarifying some issues and of stimulating the study of others.

References

1. Gulliksen, H. *Theory of Mental Tests.* New York: Wiley, 1950.
2. Thorndike, Edward L. "Reading as Reasoning: A Study of Mistakes in Paragraph Reading," *Reading Research Quarterly*, 6 (Summer 1971), 425-434.
3. Tuinman, J. Jaap "Thorndike Revisited—Some Facts," *Reading Research Quarterly*, 7 (Fall 1971), 195-202.

Robert Karlin
Queens College
City University of New York

EVALUATION FOR DIAGNOSTIC TEACHING

A good reading program is one that develops the basic skills students need in order to read, that teaches them how they can use reading as a tool for learning, that fosters an appreciation of literature, and that develops permanent interests in reading for enjoyment. These four characteristics become the objectives of our instructional program and at the same time serve as guidelines for evaluating the progress children make in reading.

Reading is not a simple skill nor even a single skill. Children do not master reading in one or two years just as they do not master any other complex activity in a brief period of time. They learn some reading skills and develop some attitudes toward reading as they complete one stage of development and move into another. What they may be able to accomplish at one point in their reading development will not be good enough at another. This fact explains why some children can cope with early reading demands but not later ones. It also underscores the need for continuous evaluation and orderly reading experiences based upon such evaluation.

We say that children learn to read; what we really mean is that children master the skills and develop the attitudes they need in order to acquire the ability to read. Children with reading ability draw upon a body of skills that they use to understand and assimilate printed messages. All children do not necessarily use the same skills in reading identical materials. Their levels of achievement and the nature of the reading task determine which ones they apply. Some children are more efficient than others in using their skills.

To fulfill the requirements of diagnostic teaching we may define the objectives of reading programs by identifying the areas on which teachers need to focus attention. The kind and amount of reading growth children achieve is proportional to the degree to which teachers manage to translate the objectives into learning tasks and guide children in mastering them. Teachers who operate within this framework will view the functions of testing much differently from those whose main concern is to grade pupils. Thus they do not ask such broad questions as how well children identify words and know their meanings, how well they understand what they read, and how well they read for information. Instead, they realize that there are more basic questions for which they must seek some answers in order to meet the requirements of diagnostic teaching: How well do pupils respond to different types of context clues? What pronunciation problems do they meet as they use the respellings of the dictionary? How well do they identify important ideas when they are stated and when they must be inferred? The answers to these and other pertinent questions help teachers decide what in the reading curriculum requires specific treatments. Moreover, this kind of evaluation suggests what types of instructional materials will be required and what their levels of difficulty should be.

Diagnostic teaching benefits children who are making satisfactory progress in reading. Teachers can anticipate superior results as they work with children who are experiencing difficulties in learning to read, if there is a positive relationship between the problems and the remedies. Inherent in the concept of diagnostic teaching is the idea that evaluation is an ongoing activity as long as instruction continues. The teacher formulates plans from the information he acquires about his pupils, but he knows that as he teaches he will receive new data. It is likely that he will have to modify his practices in order to satisfy the children's current learning needs. Occasionally he may have to revise his practices drastically.

This need for continuing evaluation raises questions about the initial effort to obtain information about children's reading. How extensive should the analysis be? Should many different tests be administered to children before instruction begins? There are differences of opinion about these matters, but it seems reasonable to suggest a middle course. Instead of spending many hours testing children's reading initially, teachers can take the time to find out

where on the reading ladder children are and what some of their reading needs appear to be. Although this information is incomplete, and possibly somewhat inaccurate, it can be used to plan early reading lessons. As teachers work with children they will confirm and revise their initial judgments and note new behaviors that affect their reading plans. These practices are so much better than hit-or-miss, trial-and-error teaching.

Teachers can appraise children's reading by using standardized tests and informal measures. Each form of appraisal provides information that may be useful in assessing what they are doing and planning suitable activities for them.

STANDARDIZED TESTS

Most schools administer survey or diagnostic-type reading tests which provide general information about students' reading and an estimate of their reading achievement. The latter tests are supposed to identify with greater precision what the reading strengths and weaknesses of students are, but such is not always the case. Although standardized reading tests suffer from a number of weaknesses, teachers can extract some useful information from their results.

Most standardized reading tests yield separate grade placement or percentile scores for each section of the test. A wise teacher will not merely be concerned with the test's total score but will want to know how it was obtained. Thus he can determine if the pupils are equally strong in all areas tested or if some pupils are stronger in one area than another. Pupils may have the same total score but obtain it in different combinations of subtest scores. This first analysis may indicate which children need help in one or more areas. A more careful examination of the composition of test items and the children's responses to them might provide the teacher with information about their specific reading requirements. Some reading tests which are presumably diagnostic identify the subskills so that teachers can categorize responses. Most reading tests are not sufficiently refined to enable teachers to make such an analysis easily, but more can be learned from the children's responses to test items than was realized in the past. One technique is to sit down with children and go over the test items with them. Perhaps they can explain how they decided on their responses. It is possible that even correct responses were

reached in inappropriate ways or that children guessed many of the answers. Teachers may be able to discern patterns of errors by comparing similar test items and responses. Standardized reading tests suffer from real weaknesses, but the effect of these weaknesses would be lessened if teachers used them with more understanding.

One caution is particularly important in using standardized tests for diagnostic teaching: the grade equivalent that a test assigns to a child's performance is usually higher than the publisher's grade level designation of appropriate instructional materials for the same child. Particularly for the child who is having reading difficulties, the most effective reading materials may be graded considerably below his grade score on the test. Furthermore, grade scores at the lower and upper ends within a range of possible scores are not as valid as those which fall in the middle. These weaknesses are due to problems of test construction and statistical treatments. Tests that cover many grades suffer more from this weakness than those intended for one or two. In addition, one must take the standard error of measurement into account when interpreting test results. It is better to think of a child's achievement as falling within a range of scores than as a single score.

We should recognize that the kinds of reading that tests require do not cover all the types of reading that children engage in. Tests do not demand the sustained reading children do in school and elsewhere. It is one thing to understand a single paragraph and another to react suitably to a longer passage. Children ordinarily do not read words in isolation nor do they have to read under timed conditions which do not allow for much flexibility. Reading tests offer approximations of how well children read; values they do not possess should not be ascribed to them.

Teachers may use standardized reading tests if they understand what their limitations are and are able to interpret their results adequately. The tests permit us to speak with some objectivity about the reading achievement of children.

TEACHER-MADE TESTS

Teachers are depending more and more on their own evaluations of children's reading. This does not mean that they merely observe children read and in haphazard fashion decide what their

reading instruction ought to be. Instead, they follow fairly well-established procedures to find out how well pupils are reading and plan their programs accordingly.

One form of teacher-made or "informal" reading test does not yield a grade-placement score, but it does help teachers identify independent, instructional, frustration, and expectancy achievement levels. Although there is not complete agreement on the standards required for each achievement level, teachers will not be too far off the mark if they adhere to a reasonable range in which they expect children to perform, as well as take into account observable reactions as children read orally and silently.

Teachers may determine from oral reading performances what problems pupils have in recognizing words. Some pupils may consistently omit certain inflectional endings, confuse vowel sounds, or fail to utilize roots in unknown words. Patterns of errors might be discernible and serve as a basis for planning lessons to overcome specific weaknesses. If silent reading is followed by suitable questions, pupils' answers will reveal not only how well they understand stated ideas but also how deeply they are able to probe ideas. These analyses would be the base for initiating instruction and continuing the study of reading needs.

A less accurate but quick way to estimate a child's reading achievement level is to have him read words on a list that samples vocabulary from a graded series of books. A separate list of words could be prepared from the vocabulary represented in readers, social studies books, and science books. The primary word lists would contain about twenty words each and the higher-level lists thirty or more. If the child missed much more than 10 percent of the words on any list, that could indicate that the materials are too difficult for him. A comparison of results from reader and subject lists could reveal differences in difficulty between the two. Children may have more trouble reading science textbooks than social studies books or readers.

Another way to estimate the difficulty children will have with materials is to apply the cloze procedure to two or three typical excerpts drawn from it. Every tenth word is removed from each excerpt and replaced by a blank of standard length; the reader is expected to supply the missing word. If he is able to supply somewhat less than half of the missing words in the excerpts, he can probably comprehend the material sufficiently well to profit from

instruction in it. [The research relating cloze scores to instructional level has been done with cloze tests in which every *fifth* word was deleted (*1*). To children, however, every fifth word cloze tests appear formidable; every tenth word cloze tests seem more appropriate for classroom use.]

A teacher may gain some insights into ways children read by studying their responses to cloze exercises. Pupils may fail to relate earlier ideas provided by the text to ones offered later, or they may become confused by certain sentence structures. Problems could surface as the teacher encourages pupils to describe how they decided what the missing words were.

Teachers can prepare individual and group tests to determine how well pupils manage specific skills. These tests should contain enough items to assure that each skill is adequately sampled. Care should be taken that the exercises require the pupils to perform the intended skill or demonstrate knowledge of it. These tests would need to be prepared in the same way as the others—sets for each achievement level.

The aim of diagnostic teaching is to identify growth areas in which children are progressing satisfactorily as well as pinpoint others to which greater attention should be given. Teaching plans are based on children's reading performances and directed toward specific learning tasks. Initial appraisal precedes instruction and reveals where children are on the reading continuum. Further evaluation accompanies instruction and provides teachers with information they need to make their teaching relevant.

Reference

1. Rankin, E. F., and Joseph W. Culhane. "Comparable Cloze and Multiple-Choice Comprehension Test Scores," ***Journal of Reading***, 13 (December 1969), 193-198.

Wayne Otto
University of Wisconsin
at Madison

EVALUATING INSTRUMENTS FOR ASSESSING NEEDS AND GROWTH IN READING

Evaluating an instrument for assessing needs and growth in reading amounts to answering two questions.

1. What do I want to know?
2. Does this instrument (or technique) do the job?

Clearly, what one wants to know will suggest the approach that must be taken; therefore, the answer to the first question sets the stage for answering the second one.

Here we shall consider three main approaches to assessment: standardized achievement tests, criterion-referenced measures, and informal procedures. Means for estimating pupils' capacity, although they are important in assessing needs and growth, will not be considered; the discussion is limited to the assessment of reading behavior.

STANDARDIZED ACHIEVEMENT TESTS

In general, the standardized achievement tests are norm referenced. That is, a given individual's performance is examined in relation to the performance of other individuals. The following points should be considered in choosing standardized tests.

1. ***Define the purpose for testing.*** Standardized tests may be given for any number of reasons—to compare class achievement with local or national norms, to determine the current achievement status of classes or individuals in order to learn whether corrective or remedial steps should be taken, to screen in order to determine the need for further testing, or to evaluate the developmental program. When the purpose for testing is clearly in mind, a

decision can be made as to whether a *survey test* or an *analytical test* would best suit the purpose. Generally, survey tests are group tests designed to provide a score that will tell the teacher how well a class or a pupil compares with other pupils of the same age and grade. Survey tests are typically used at the survey level of diagnosis. Analytical tests may be either group or individual tests. They are designed to break down the total reading performance into specific strengths or weaknesses. Group tests have the obvious advantage of testing more pupils in less time than individual tests, but the latter are likely to provide much more information regarding the idiosyncracies of an individual's performance.

2. *Locate suitable tests.* From among the many tests currently available, several will typically appear to be appropriate for the purpose identified. Probably the most useful single source of assistance in locating and sorting out suitable tests is *The Sixth Mental Measurements Yearbook*, edited by Oscar K. Buros. (Previous editions were published in 1938, 1940, 1949, 1953, and 1959.) Available tests in education and psychology are listed and described in the yearbook. Brief descriptions of such things as cost, coverage, and source, as well as one or more critical reviews, are included for each test.

3. *Evaluate before selecting.* Once the tests that appear to meet the requirements of a given situation have been identified, they should be carefully evaluated in terms of such things as reliability, validity, economy, ease of administration, adequacy of the manual, relevance of the norms provided, and appropriateness of the content for local pupils.* A test that is *reliable* yields consistent results. A test that is *valid* actually measures what it is supposed to measure. The validity of a test can be estimated by correlating individual scores on the test with performance on a previously selected criterion task or test. The fact remains, of course, that many highly regarded, widely used tests have only face validity. That is, they appear to measure what they are intended to measure.

An adequate test manual includes the following kinds of information: 1) Clear and concise directions for administering the test. This is important because a major reason for using a standardized

*For an extended discussion of things to consider in selecting tests consult *Standards for Educational and Psychological Tests and Manuals*, American Psychological Association, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

test is to secure data under stated conditions. 2) Adequate information regarding the reliability and validity of the test. 3) Norms based upon sound sampling procedures. That is, the sampling of scores upon which the norms are based should be large and distributed according to geographic location and socioeconomic areas. 4) Aids for interpretation. Provision for profile analysis and illustrative interpretations are useful.

Finally, a test must be readily and currently available if it is to be used in quantity. It should be economical: such things as initial cost of test booklets, whether the booklets are reusable, ease of scoring, and compatibility with machine scoring techniques must be considered. A test that is very reasonable in terms of initial cost could be prohibitively expensive in terms of time required for scoring or replacement costs. Availability of alternate forms is required if the test is to be used in test-retest comparisons.

The best way to become completely familiar with a test is to take the test yourself and then to administer it to a few children. There is no better way to learn about problems in administration and scoring and the appropriateness of the content. Specimen sets of tests are readily available from publishers at a reasonable cost.

LIMITATIONS OF STANDARDIZED TESTS

Standardized tests share some rather severe limitations that ought to be kept in mind even after the "best available" test is chosen. Some of the more salient limitations are given here.

1. The very fact that a test is "standardized" in terms of administration and scoring may make it inappropriate for use with certain groups or individuals. The test may be too difficult or too easy; items may be meaningless or placed at inappropriate levels; directions may be incomprehensible.

2. The test maker's quest for brevity, which unfortunately but pragmatically enhances the salability of tests in some circles, may result in unrealistic time limits and a choice between depth and breadth in sampling. Scores of children who work very slowly but accurately are likely to be meaningless; the sampling of behavior is likely to be superficial or constricted.

3. Group administration may work to the disadvantage of certain individuals. The group situation combined with the standardized conditions may invalidate the test in some instances. For example, a child who fails to understand one or two words in a set

of directions may be unable to respond to any of the items, which he may or may not have known.

4. The format of the test may restrict the type of items used. A machine scorable format, for example, virtually demands some form of multiple-choice items. Certain behaviors are not adequately sampled with multiple-choice items.

5. Tests at upper grade levels assume ability at lower levels. Thus, a pupil may be able to score at a certain base level by simply signing his name to the test booklet. Furthermore, it is generally acknowledged that standardized tests do tend to yield overestimates of appropriate instructional level.

CRITERION-REFERENCED MEASURES

Criterion-referenced measurement relates test performance to absolute standards, usually stated in terms of behavioral objectives, rather than to the performance of other students. Such measures are most useful for assessing pupils' mastery of specified objectives. Some of the salient contrasts between norm-referenced (standardized achievement) tests and criterion-referenced measures follow.

1. Standardized tests have a low degree of overlap with the objectives of instruction at any given time and place. The overlap for criterion-referenced measures is absolute, for the objectives of instruction *are* the referents.
2. Norm-referenced tests are not very useful as aids in planning instruction because of the low overlap just mentioned. Criterion-referenced measures can be used directly to assess the strengths and weaknesses of individuals with regard to instructional objectives.
3. Again because of their nonspecificity, norm-referenced tests often require skills or aptitudes that may be influenced only to a limited extent by experiences in the classroom. This cannot be so for criterion-referenced measures because the referent for each test is also the referent for instruction.
4. Standardized tests do not indicate the extent to which individuals or groups of students have mastered the spectrum of instructional objectives. Again, there is no such problem with criterion-referenced measures because they

focus on the spectrum of instructional objectives in a given situation.

The main advantage of criterion-referenced measures is that they get directly at the performance of individuals with regard to specified instructional objectives. The sensible management of a system of individualized instruction requires knowledge of each pupil's performance with regard to the objectives of the system. Such knowledge can be derived from criterion-referenced measures.

LIMITATIONS OF CRITERION-REFERENCED MEASURES

There are of course some dangers and pitfalls in criterion-referenced measurement.

1. Objectives involving hard-to-measure qualities, such as appreciation or attitudes, may be slighted.
2. Objectives involving the retention and transfer of what is learned may become secondary to the one-time demonstration of mastery of stated objectives.
3. Specifying the universe of tasks (determining critical instructional objectives) to be dealt with is of extreme importance. Good tests will do nothing to overcome the problem of bad objectives. But note that the problem here is no different for norm-referenced testing.
4. Determining proficiency standards can be troublesome. Perfect or near-perfect performance should be required if a) the criterion objective calls for mastery, b) the skill is important for future learning, c) items are objective type and guessing is likely. Less demanding performance may be adequate if any of the three conditions do not prevail.

Fortunately, one does not need to choose between norm-referenced and criterion-referenced measures. To the contrary, the two types of measures ought to complement each other, with each type chosen according to the purpose for testing.

At the present time the biggest problem with criterion-referenced testing may be to find such tests to consider. Certainly the movement toward criterion-referenced testing is in its infancy compared to norm-referenced testing. As a consequence, once one has decided to take an objectives-centered approach to instruction, he may be confronted with the task of devising his own criterion-referenced measures as discussed in the section that follows. While

such a task may at first seem overwhelming, it may turn out to be a good thing if it causes one to look a bit more carefully at what one is doing and how to assess it.

INFORMAL PROCEDURES

In the process of diagnosis a teacher will often find it necessary to seek information that is not available from existing tests or to supplement information from them. When this is so, it is up to him to devise his own informal measuring device. Since in many instances the teacher will want to know whether particular students know how to do a particular task, the measurement, though informal, is likely to be criterion-referenced. The following sequence can serve as a guide to the effective use of informal assessment: First, decide exactly what information is desired and what this means in terms of observable behavior; then devise new or adapt existing test items, materials, or situations to sample the behavior to be evaluated; keep a record of the behavior evoked in the test situation; analyze the information obtained; and finally, make a judgment as to how the information fits the total picture and how well it fills the gap for which it was intended.

Examples of some of the most useful and most used informal devices for gathering diagnostic information, particularly regarding strengths and weaknesses in specific skill development, follow:

1. ***Informal observation***. The most naturalistic informal technique for gathering diagnostic information is informal observation of the pupil. This technique is often overlooked; but it is one that alert, skillful teachers can use effectively for a number of purposes—systematically observing a child's overall performance, learning about his interests and attitudes, finding out about his approaches to problem solving and to study situations, and detecting physical problems and limitations. Observing with a purpose can provide the teacher with real insight into the problems a child may be encountering when he attempts to follow through story problems in arithmetic, attack new words, or write legibly.

2. ***Anecdotal records***. In its simplest form, an anecdotal record can consist of a manila folder in which word samples and observations are kept in chronological order. The primary purpose for keeping such a record is to help the teacher keep in mind the developing characteristics of a child. Gradual but steady improvement may be seen as lack of improvement if there are no readily

available checkpoints. Obviously, the record loses its value if it is simply cluttered with an occasional drawing and general statements like, "Clyde appears to be doing better." Entries must be dated.

3. *Informal tests*. Many of the books, workbooks, and periodicals designed for school use include informal, nonstandardized tests that can be used for quick checks of pupils' comprehension, writing ability, grasp of arithmetic concepts, and the like. Similar informal tests can be constructed by the teacher to check on pupils' grasp of just-presented material or to get samplings of various kinds of behavior.

4. *Checklists*. In this general category are included such things as interest and personality inventories; questionnaires of work habits, interests, activities, associates; and lists of specific skills that can be used to check a pupil's mastery of certain areas. The lists are a practical means for systematizing observations.

5. *Informal reading inventories*. In the area of reading, many teachers use an informal reading inventory to observe a pupil's oral and silent reading at several difficulty levels. The inventory consists of samples from the various grade levels of a basal reader series plus comprehension questions. Four levels of reading ability are typically identified through the use of the inventory: a) independent level—the level at which the pupil can read independently with at least 99 percent accuracy in word recognition and 90 percent or better comprehension; b) instructional level—the level at which the pupil can read with some help from the teacher; c) frustration level—the level at which the pupil can no longer function effectively; and d) listening capacity level—the highest level at which the pupil can comprehend at least 75 percent of material that is read to him.

Each of the informal devices listed can be adapted in a number of ways to increase its applicability. Note that all of the informal procedures discussed lend themselves very well to criterion-referenced measurement. Once criterion behaviors have been identified, they can be sampled with paper-and-pencil tests or through informal procedures.

Dale D. Johnson
University of Wisconsin
at Madison

GUIDELINES FOR EVALUATING WORD ATTACK SKILLS IN THE PRIMARY GRADES

The terms *word attack, word analysis, word recognition*, and *decoding* are often used synonymously in reference to a cluster of rather diverse skills that readers employ to identify words they do not recognize in print. Six or seven skills are commonly described in reading methods textbooks within their chapters on word attack. These skills are: configuration (sight words), picture clues, phonics, syllabication, structural analysis, context, and use of the dictionary.

In the writer's opinion, only three of these usual six or seven skills are truly word attack skills that are useful to children. Considerable evidence shows that instruction in configuration—word shape—(drawing little boxes around words) is probably a waste of everyone's time (*2*). The procedure is rarely used beyond the initial weeks of first grade, and even then does little to help a child form generalizations that will be useful later. Likewise, dictionary skills—worthwhile as they are—should be treated as reference, not word attack skills. Picture clues are in essence no more than context clues, whereby information is gleaned from graphic or pictorial context, rather than through syntactic or semantic clues. Finally, it seems to me that syllabication should be treated as a subdivision of phonics when pronunciation generalizations are used, and as the basis for structural analysis when morphemic clues are used. The key word attack skills, then, and those that will be the basis for the remaining discussion, are *phonics, structural analysis*, and *context*.

Research has shown that children typically enter their first

grade classrooms with an oral/aural vocabulary of several thousand words. On the other hand, most children entering a beginning reading program cannot read more than a handful of words. Thus, the purpose of instruction in word attack is clear. It is based on the assumption that many words that are unfamiliar to a child in print are, nevertheless, words that he can use or understand in conversation. Therefore, word attack skills should enable a child to bridge the gap from unfamiliarity in print to the meaning that he already attaches to a word in his listening and speaking vocabulary. This purpose is particularly true of two word attack skills: phonics and structural analysis. The third skill, use of contextual clues, is often a vocabulary building skill as well.

The overall goal of facility in word attack is an ever enlarging sight word vocabulary—a vocabulary of words recognized instantly in print. Smith (3) estimates that adult readers have a sight vocabulary of between 20,000 and 100,000 words. Obviously, mature readers did not learn each of these as a distinct sight word. Rather, use was made of a variety of word attack skills. Teachers of reading are challenged with helping children develop those skills that they will use to increase their sight vocabularies from a few words to tens of thousands.

With this rationale for teaching word attack, the remainder of present discussion will be directed to the assessment of children's acquisition of word attack skills. Four guidelines for evaluating word attack ability will be presented. These guidelines are based on two beliefs: 1) Skill in word attack is essential for developing readers and 2) the key word attack skills can be measured.

Four guidelines for evaluating word attack will be discussed:

1. Skill in word attack should be measured through teacher-made or published tests that use synthetic (or nonsense) words.
2. Skill in word attack can be adequately measured through group-administered tests.
3. Word attack tests should measure decoding not encoding skills.
4. Word attack skills should be evaluated often in the primary grades so that programs can be geared to the needs of pupils.

USE OF SYNTHETIC WORDS

I have suggested that word attack tests should use synthetic words. The rationale for this view is that unless synthetic words are used, teachers can never be sure whether they are measuring the specific word attack skill in question or are simply measuring words that may be in a child's sight vocabulary.

The purpose of phonics, very simply, is to help a child pronounce a word he doesn't recognize in print, with the reasonable assumption that once pronounced, the word may be recognized from the child's oral/aural vocabulary. How can phonics generalizations be tested using synthetic words? If we are evaluating children's use of the "hard and soft *c* generalization," for example, it seems much more reasonable to use synthetic words such as ***ceb, cack, cobe,*** and ***cipe*** than such words as ***cent, cat, coat,*** and ***city***. If the child pronounces the latter four correctly, can we really be sure he has mastered the *c* generalization?

In terms of structural analysis, such synthetic words as ***ungate***, ***meatness***, and ***footbank*** will more accurately assess a child's knowledge and use of prefixes, suffixes, and rootwords, than would real words such as ***unhappy, happiness***, or ***stoplight***. We can be sure children have not seen the synthetic words—thus must attend to base words and affixes—whereas with the real words we may merely be determining whether or not these words are within the child's sight vocabulary.

Synthetic words are also useful when measuring the ability to use contextual clues. For example, the use of the word ***cromp*** in the sentence, "He received a new ***cromp*** with silver handlebars, purple pedals, and a chrome-plated chain," can provide a good indication of the child's attention to context in understanding an unfamiliar word. Again, where a real word (***bicycle***) is used, uncertainty arises as to what is being measured.

I tend to favor teacher-made tests of word analysis, rather than published tests, because the tests can be constructed to measure the specific word attack skills of interest. If we are interested in evaluating our success in teaching a particular skill, national norms are not needed. Tests can sometimes be very short, containing four or five items to assess a generalization or subskill.

GROUP ADMINISTERED TESTS

Educators are well aware of the many advantages of individ-

ually administered tests and also know their principal shortcoming—they take time. I would rather see teachers spending their time on instruction than on measurement. But, the instruction must be based on diagnosis. Word attack tests in phonics, structural analysis, and context clues lend themselves particularly well to group procedures.

With a small corpus of sight words to be used as distractors, all important phonics generalizations can be assessed with multiple-choice tests. A few sample items constructed to measure long and short vowel generalizations might look like the following:

"Circle the real word whose underlined letter sounds the same as the underlined letter in the word at the left."

bamp	apple	game	dare
dape	dare	game	apple
rad	apple	dare	game
bame	game	apple	dare

To assess the child's use of structural analysis, children can be asked to divide between prefixes, base words, and suffixes in such words as the following:

prehead ***doorest*** ***eatroom***

Multiple choice items can be used to evaluate the child's ability to use context in defining an unknown (synthetic) word. For example, in the sentence used earlier, the word in question was ***cromp***. Children could be asked: "A ***cromp*** is a . . . a. bird
b. bicycle
c. teacher

Once a group test of the particular word attack skills of interest to the teacher has been constructed, administered, and scored, the teacher may want to test a few children individually—those who seem to have had the greatest difficulties. But a great deal of diagnostic information can be gained, and time saved, through using group administered tests.

DECODING, NOT ENCODING

One of the major problems with many word attack tests, particularly tests of phonics ability, is that they involve encoding,

or spelling, rather than decoding, or pronouncing. The sets of grapheme-phoneme correspondences for encoding and decoding are often quite different. For example, if you were asked to write /Kɔθ/ on your paper, you would have two choices for the initial consonant, *coth* or *koth* (*coth* would be proper because /K/ is spelled *c* in initial position except in borrowed words) and several choices for the medial vowel, *coth* (scoff), *cauth* (cause), or *coath* (broad). On the other hand, if you were shown the word *coth* and asked to pronounce it, other choices are available: /kaθ/ (mop), /Koθ/ (both), /Kɔθ/ (moth), or /KʌθI (mother). The point is clear: decoding and encoding correspondences are not always bidirectional. Therefore, tests in which the examiner reads a synthetic word and the children are asked to respond, either from among choices or in writing, are not accurately measuring word attack *decoding* skills. The example items presented on the preceding page *are* based on decoding and should be the type used in reading. Read (*1*) found that young children could write (encode) a number of words but later could not pronounce (decode) their own spellings.

There is nothing wrong with testing encoding if one is interested in spelling ability, but testing should fit the purpose of the instruction. To measure childrens' ability to use phonics generalizations in pronouncing unfamiliar printed words, tests requiring decoding should be used. Failure to do so could cause teachers to plan instructional programs which do not match the skill needs of their pupils.

FREQUENT EVALUATION

Last, but certainly not least, it seems imperative that children's progress in the development of word attack skill be evaluated regularly and frequently. Surely most primary grade children will experience word attack instruction every week. Beginning early in grade one, instruction in invariant consonants and regular vowel patterns will be underway. Later, generalizations regarding variant consonants, consonant clusters, vowel clusters, and syllabication will be introduced. By second grade, children should be developing their use of structural analysis and contextual analysis. For a sound developmental reading program to flourish, it will be essential that word attack skills be assessed often. Ideally, assessment should be done after each phonics generalization, structural analy-

sis clue (base words and affixes), and contextual strategy (pictorial, semantic, and syntactic) has been taught. Frequent, planned assessment of the word attack skills will enable the teacher to design needed instructional activities geared to the individual characteristics of the pupils.

SUMMARY

These four guidelines— 1) use synthetic words, 2) construct group tests, 3) test decoding not encoding abilities, and 4) evaluate frequently— should provide a framework for continuing evaluation of the essential word attack skills. Word attack tests should be geared to the specific skills being taught. Word attack tests are not difficult to construct and can provide valuable information about the degree of success the word attack program is having. Testing should be done often, at least weekly

Word attack skills are essential for children developing their reading ability. The wise teacher will continue to evaluate these skills so that the instructional program can be most fruitful. It is hoped that the guidelines suggested here will contribute to that end.

References

1. Read, Charles. "Pre-School Children's Knowledge of English Phonology," *Harvard Educational Review*, 41 (February 1971), 1-34.
2. Samuels, Jay S. "Modes of Word Recognition," in Harry Singer and Robert B. Ruddell (Eds.), *Theoretical Models and Processes of Reading*. Newark, Delaware: International Reading Association, 1970, 23-37.
3. Smith, Frank. *Understanding Reading*. New York: Holt, Rinehart and Winston, 1971, 146-148.

Paul Conrad Berg
University of South Carolina

EVALUATING READING ABILITIES

Professionals in measurement tell us that if a thing exists, it can be measured. Reading specialists have been measuring and evaluating bits and pieces of reading abilities ever since William S. Gray first published The Gray Standardized Oral Reading Paragraphs in 1915. Also in 1915, Starch reported a silent reading test that he had devised, and with it he postulated the chief elements of reading to be comprehension, speed, and pronunciation (*19*). By 1921, Gray had effectively stated the case for silent reading, and many silent reading tests were being published, using Starch's postulated factors. Thorndike, in 1917, added the first recorded study of reading as reasoning to this growing area for research (*21*).

The idea of separate, definable skills grew so rapidly that by 1945, Burkart (*4*) reported that her survey of the literature on reading instruction indicated that "Reading is not a single act but a complex activity made up of at least 214 separate abilities. These abilities are motor, sensory, or intellectual in nature."

Today, some fifty years after the first studies in evaluation by Gray and Starch, a review of reading tests turns up well over 70 reading abilities that publishers tacitly infer are defensible as separate factors. Buros' *Reading Tests and Reviews* (*5*), published in 1968, contains some 500 pages devoted to the task of describing and evaluating published reading tests, and the latest edition of Buros' *The Seventh Mental Measurements Yearbook* (*6*) adds a hundred more pages of technical data on reading tests.

The purpose behind testing and evaluation was (and is) not simply to list the reading abilities of the single student or groups

of readers. Through testing, remedial reading techniques developed and individualized instruction came into being. The overall improvement in the teaching of reading also was promoted by this evaluation movement.

With literally dozens of reading skills hypothesized as measurable, we have seen an astonishing outpouring of workbooks, kits, and visual aids of every description that purport to teach the skills as measured by the reading tests. But what if the separate skills that we claim to teach, such as retention of details, ability to determine the intent of the writer, ability to grasp the general idea, and on *ad infinitum* do not in reality exist as separate measurable factors, at least as measured by our present reading tests? What would this suggest, in effect, about the materials that are specifically created to improve these separate skills?

Before a reading test can claim to measure some particular ability in reading, the existence of that ability must first be demonstrated by an appropriate statistical analysis. What have such analyses indicated? One of the first attempts at a factor analysis of reading comprehension ability was made by Traxler (*23*) in 1941. He sought to discover whether or not the separate parts of the VanWagenen-Dvorak Diagnostic Examination of Silent Reading Abilities did indeed yield "measures which are independent enough to warrant their separate measurement and use as a basis for diagnostic and remedial work." After administering the test to 116 tenth grade students, Traxler stated that the five sections of the test appeared to be measuring the same abilities, and doubted that the separate scores contributed anything over the total reading level score.

Also in 1944, a factorial study of reading abilities was made by Davis (*7*). Of the nine variables hypothesized by Davis, five were found to meet his criteria for stability and order as variables. These factors were knowledge of word meanings, verbal reasoning, sensitivity to implications, following the structure of a passage, and recognizing the literary techniques of the writer. In 1946 Thurstone (*22*) reanalyzed Davis' data and concluded that Davis had no statistical ground for his claim, but that there was only a single general factor comprising reading ability. Hall and Robinson (*10*), in a 1945 factorial study, identified attitude of comprehension accuracy, rate of inductive reading, word meaning, rate for reading unrelated facts, and chart reading skill as separate factors.

Stolurow and Newman (*20*) in 1959 identified only semantic difficulty (words) and syntactical difficulty (sentences) as factors determining the reading difficulty level of passages.

Hunt (*11*) and Alshan (*1*) also attempted factor analyses of reading comprehension, using test items from the Davis studies to build their research instruments. Hunt concluded from his results that only two skills in reading comprehension were factorially defensible—word knowledge and paragraph comprehension. Alshan was also unable to substantiate the hypothesis that five different skills of comprehension were independently factorable from the Davis iter .

Later, Davis (*8*) again attempted a statistical analysis of reading comprehension, using a much improved technique of cross-validation uniqueness analysis based on a sample of 1,100 high school seniors in the Philadelphia area. Eight separate skills were selected for study after an analysis of previous research, including the Davis, Hunt, and Alshan studies. Five skills were found to show a significant degree of independence. The skill making up the largest percentage of variance was "memory for word meanings" with 32 percent of the total variance of the eight variables. Next, in order, were drawing inferences from content (20 percent of the variance); following the structure of a passage (14 percent of the variance); recognizing a writer's purpose, attitude tone, and mood (11 percent of the variance); and finding answers to questions asked explicitly or in paraphrase (10 percent of the variance).

In 1969, Schreiner, Hieronymus, and Forsyth (*18*) reported a carefully conducted experiment on the reading comprehension of fifth grade pupils in nine Iowa public schools. The purpose of the study was to provide classroom teachers with information relative to what traits of comprehension are measurable so that useful diagnostic tests could be provided. The eight factors investigated were: speed of noting details, speed of reading, paragraph meaning, determining cause and effect, reading for inferences, selecting the main idea, verbal reasoning, and listening comprehension. Only four factors were found to be statistically definable for diagnostic purposes: speed of reading, listening comprehension, verbal reasoning (classification of words), and speed of noting details.

Benz and Rosenier (*3*) in 1968 reported a study that attempted to equate certain word analysis skills with comprehension. Using the Gates Level of Comprehension Test as the criterion

and the subtests of the Bond, Clymer, and Hoyt Silent Reading Diagnostic Tests as the predictor variables, they found that the subtests having the greatest relationship to the criterion were *words in context, rhyming sounds*, and *syllabication*. Subtests having low statistical relationship to comprehension were the *root word, word elements,* and *beginning sounds.*

There are many more studies in the literature that add to the same generalization: there are few consistent findings relative to a large number of statistically identifiable separate reading abilities. This review also suggests that research in measurement and classroom practice in measurement have little in common. If one were to take a rough average of the number of factors that researchers suggest can be measured independently, one would come up with a number somewhere between two and five. Lennon (*14*), writing on the same subject in 1962, suggested that only four factors could be measured reliably: 1) a general verbal factor, 2) comprehension of explicitly stated material, 3) comprehension of implicit or latent meaning, and 4) appreciation. While several studies subsequent to Lennon's review have been discussed, four factors might still be close to the number measurable. Yet, as already stated, a review of reading tests turns up 70 or 80 factors that various tests implicitly claim to measure.

Even though the results from standardized tests of comprehension may not measure separately the several skills they claim to measure, it is possible that from a pragmatic or practical point of view such tests may have some value for reading instruction. Obviously, teachers do not teach "pure" skills in isolation any more than tests can measure them. Therefore, it is possible that the teacher who gives the tests gets from the data the kind of information that is needed to improve instruction, even though neither the test results nor his teaching deal with precisely defined or measured skills. The question of evaluation effectiveness is really meaningless, however, unless we see what effect evaluation has on the instructional materials and practices that are, in part at least, an outcome of differential testing. That is, does testing make a difference for instruction in any way, either by making the materials more focussed and effective or by significantly changing instruction?

It seems evident that the same subjective rationalization that helped produce our measurement techniques is also responsible for

the methods and materials that we use in teaching reading comprehension. Some researchers have stated that methods and materials for teaching reading are no more scientific than the *a priori* pronouncements prior to the research of this period. For example, in 1941, Laycock and Russell (*13*) reported that an analysis of reading improvement manuals revealed that few of them had any basis in research findings for their suggested reading improvement. While the findings of this early study are perhaps not surprising, in 1950 Robinson (*16*) made the same charge when he stated "no particular professional acuity is required to penetrate the superficiality of the types of exercises and treatment that characterize most of these volumes." Atwater (*2*) in 1968, studied eleven popular reading improvement workbooks used at the college level to determine if the skills they claimed to improve were actually defined, and secondly, what aspects of the defined skills were actually taught and measured in the workbook exercises. He found, for example, that definitions of comprehension covered the range of ambiguity from simply the word "understanding" in one workbook, to "an ability to grasp the author's thought structure as an organized whole" in another. From 80 to 95 percent of the exercises in the workbooks, including questions, dealt with factual, detailed information. One workbook, for example, claimed that comprehension included knowledge of structure and style of writing. The one question in the workbook that was meant to measure this skill was "how many paragraphs are in the preceding article?"

And yet teaching, to be successful, must be directed. A teacher must know what his pupils can and cannot do in terms of common behavioral objectives in reading. Overreliance on published tests can create a false sense of having information that indeed one does not have. Evaluation is much more than testing—it must include a variety of observations. One important type of observation is guided by teachers' daily questions. Skillful questioning by teachers is not only an art of evaluation, but also a part of good teaching. A facet of this function is learning the art of asking skillful questions and leading the student to develop a questioning attitude about everything that he does. How skilled are teachers in this characteristic? Floyd (*9*) studied the verbal activity in the classes of 40 teachers selected from administrative ratings as the best teachers in a city school system. He recorded a significant amount of verbal activity in these teacher's classes and separated

out for analysis that verbal activity dealing with teacher-student questions. Of all the questions asked, 96 percent were asked by the teachers, only 4 percent by students. Only 5 percent of the teachers' questions—one in 20—demanded a thought answer or seemed capable of creating any stimulation or reflection on the part of the student. Eighty-five percent of the teacher-initiated questions fell into only two categories—memory for facts and information. Almost all could be answered by "yes" or "no" or simple repetition of a stated fact. Questions dealing with problem solving, the student's interests, or for helping to locate student problem areas in learning were almost never asked. In another study (*15*), 190 teachers were asked to list as many reasons or purposes as they could think of for asking questions. Fewer than three reasons were given per teacher for asking questions. Only 19 of the 190 went beyond the need to ask simple, factual questions. Only 10 percent said that teacher questions should require pupils to use their facts to make generalizations and inferences. Satlow (*17*), by comparison, lists 120 reasons for asking questions. These are just a few of his suggestions: "Do you challenge students by asking questions that arouse their curiosity for further knowledge? Do your questions stimulate thinking on the part of students and help to develop in them effective methods of attack? Do they help guide wholesome interaction among the students? Do your questions disclose the degree to which a spirit of inquiry has been established? Do they place the burden of thinking on the students?"

SUMMARY

Knowing about how our students learn is more than an evaluation of a compilation of scores from a series of standardized tests. Such tests do give us information for instruction that would be difficult or time consuming to get otherwise. But to complete the picture of a student's learning pattern, become skilled in the informal, observational inventory technique that makes you a diagnostic teacher—the best kind that there is. Kress (*12*) summarizes this conclusion, describing daily, diagnostic teaching techniques under the headings of general observation, observation of listening situations, speaking situations, and reading situations. As the child listens, for example, can he follow directions? Can he "picture"

things described in words? In speaking situations does he use past experiences, logical argument, supporting evidence? Is he consistent or inconsistent? In reading situations, there are a multitude of observations that can be made, such as, can he find a fact or idea by skimming? Does he try to size up organization? Does he use aids, such as graphs and charts?

And so we have made the full circle, back to the teacher as the one who makes the difference. The summation of excellence has not changed for two thousand years. Tests and materials cannot duplicate teacher excellence or substitute for it. Through it the human equation remains the master.

References

1. Alshan, L. M. "A Factor-Analytic Study of the Items Used in the Measurement of Some Fundamental Factors of Reading Comprehension," unpublished doctoral dissertation, Teachers College, Columbia University, 1964.
2. Atwater, J. "Toward Meaningful Measurement," ***Journal of Reading***, 11 (March 1968), 429-434.
3. Benz, Donald A., and Robert A. Rosemier. "Word Analysis and Comprehension," ***Reading Teacher***, 21 (March 1968), 558-563.
4. Burkart, Kathryn H. "An Analysis of Reading Abilities," ***Journal of Educational Research***, 38 (February 1945), 430-439.
5. Buros, Oscar K. (Ed.). ***Reading Tests and Reviews***. Highland Park, New Jersey: Gryphon Press, 1968.
6. Buros, Oscar K. (Ed.). ***The Seventh Mental Measurements Yearbook***. Highland Park, New Jersey: Gryphon Press, 1972.
7. Davis, Frederick B. "Fundamental Factors of Comprehension in Reading," ***Psychometrika***, 9 (September 1944), 185-197.
8. Davis, Frederick B. "Research in Comprehension in Reading," ***Reading Research Quarterly***, 3 (Summer 1968), 499-545.
9. Floyd, William D. "Do Teachers Talk Too Much?" ***Instructor***, 78 (October 1968), 53.
10. Hall, W. E., and F. P. Robinson. "An Analytical Approach to the Study of Reading Skills," ***Journal of Educational Psychology***, 36 (October 1945), 429-442.
11. Hunt, Lyman C., Jr. "Can We Measure Specific Factors Associated With Reading Comprehension?" ***Journal of Educational Research***, 51 (November 1957), 161-171.
12. Kress, Roy A. "Classroom Diagnosis of Comprehension Abilities," Conference on Reading, University of Pittsburgh Report, 22 (1966), 33-41.
13. Laycock, Samuel R., and David H. Russell. "An Analysis of Thirty-Eight How-to-Study Manuals," ***School Review***, 49 (May 1941), 370-379.

14. Lennon, Roger. "What Can Be Measured?" ***Reading Teacher***, 15 (March 1962), 326-337.
15. Pate, Robert T., and Neville H. Bremer. "Guiding Learning Through Skillful Questioning," ***Elementary School Journal***, 67 (May 1967), 417-422.
16. Robinson, H. Alan. "A Note on the Evaluation of College Remedial Reading Courses," ***Journal of Educational Psychology***, 41 (February 1950), 83-96.
17. Satlow, David. "120 Questions About Your Questioning Technique," ***Business Education World***, 49 (February 1969), 20-22.
18. Schreiner, Robert L., A. N. Hieronymus, and Robert Forsyth. "Differential Measurement of Reading Abilities at the Elementary School Level," ***Reading Research Quarterly***, 5 (Fall 1969), 84-99.
19. Smith, Nila Banton. ***American Reading Instruction***. Newark, Delaware: International Reading Association, 1965.
20. Stolurow, L. M., and R. J. Newman. "A Factorial Analysis of Objective Features of Printed Language Presumably Related to Reading Difficulty," ***Journal of Educational Research***, 52 (March 1959) 243-251.
21. Thorndike, Edward L. "Reading as Reasoning: A Study of Mistakes in Paragraph Reading," ***Journal of Educational Psychology***, 8 (June 1917), 323-333.
22. Thurstone, L. L. "Note on a Reanalysis of Davis' Reading Tests," ***Psychometrika***, 11 (September 1946), 185-188.
23. Traxler, Arthur E. "A Study of the VanWagenen-Dvorak Diagnostic Examination of Silent Reading Abilities," ***Educational Records Bulletin***, No. 31 (January 1941). New York: Educational Records Bureau, 33-41.

Walter H. MacGinitie
Teachers College
Columbia University

WHAT ARE WE TESTING?*

Standardized reading achievement tests usually consist of at least two subtests—a vocabulary subtest and a comprehension subtest. Other subtests are also often included—for example, a test of reading speed. Or the vocabulary test may be subdivided into two different types of vocabulary tests, or the comprehension subtest may be divided into two or more different types of comprehension tests. What is it that is being tested by these vocabulary and comprehension subtests and by the further breakdown of vocabulary or comprehension?

The first point is that there is as much of a difference between different educational levels of the same subtest as there is between subtests with different names at the same level. The great changes that take place in ***arithmetic*** achievement tests from one grade to another are self-evident to most people. To score well on an arithmetic test for the sixth grade, a student must know a lot of things about decimals and fractions that have no bearing on performance on a test for the second grade. Most teachers and researchers are now also aware that what is measured by so-called intelligence tests changes considerably from the infant level to the intermediate grades. In contrast, the rather large change in the content of reading tests from the first to the later grades is frequently not taken into account. Although most people readily see or already recognize the different requirements posed by reading tests at different grade levels, they seem seldom to consider these differences

*Adapted from a paper presented to the Thirty-Fifth Annual Conference of the Educational Records Bureau, New York, October 1970.

when interpreting research findings or a child's educational status.

Grade changes in reading tests are most obvious in the vocabulary subtest. The easiest items for the first grade usually use simple words well known to all children in speech. The distractors, or wrong answers from which the children may choose, may all look and sound quite different from the right answer and be quite unrelated in meaning. In slightly harder questions, the distractors will present possible perceptual confusions, so that if the right answer is ***house***, distractors might be ***horse*** or ***mouse***. The vocabulary questions gradually are made more difficult by using words that are less likely to be known as sight words or words that include more difficult letter combinations.

Eventually, as the items get more difficult, the main difficulty for most children comes from uncertainty about the meaning of the words. The majority of the older children can puzzle out the pronunciation of most of the words whose meanings they know. They can even give a reasonable pronunciation to nonsense words. The test maker simply runs out of meaningful possibilities for making items more difficult by means of perceptual similarities alone. But we recognize that, for an older child, having a good reading vocabulary means more than just being able to pronounce words. The developing student learns new word meanings that a few years ago were not familiar to him in speech. Some of these new words may even now be unfamiliar to him in speech, but their meaning is recognized in print. Thus, a reading vocabulary test for older children is more concerned with whether the child understands a variety of words that he may find in written material.

This change occurs gradually in tests intended for increasingly more able readers. The title of the test remains the same ("reading vocabulary" or whatever the testmaker chooses to call it), but the ability that is tested appears to change quite radically. As represented by the harder items in a third grade test, or by the majority of items on a fourth grade test, the reading vocabulary test has evolved into a test that is nearly indistinguishable from the vocabulary section of many group intelligence tests. Thus, the correlation between a reading vocabulary subtest at the fourth grade level and a verbal IQ test is likely to be as high as the correlation between the reading vocabulary subtest and a reading comprehension subtest.

Grade changes in reading comprehension tests roughly parallel

those described for reading vocabulary subtests, though they are perhaps less drastic and less obvious. In the primary grades, the comprehension tests are more concerned with the straightforward interpretation of concrete statements and relationships, often those that are easily pictured. Sentences are simple, the number of items to be related is limited, and items to be related are not widely separated in the text. In later grades, greater stress is laid on inferences, on understanding complex ideas and difficult sentences, and on applying background knowledge.

Since these grade changes in reading tests are so obvious—particularly in the case of the vocabulary subtest—why aren't they more prominent in our thinking about the meaning of reading test scores? I believe there are at least two reasons. We recognize the changes in the content of arithmetic tests partly because these changes reflect the formal introduction of specific topics in our teaching of arithmetic. We introduce long division or the addition of fractions as a specific topic of instruction. We don't expect the students to know much about these operations before they are formally taught and, after they are taught, we expect to see them featured in arithmetic achievement tests. Except for the so-called decoding stage of reading instruction, we don't have such clear-cut ideas about separate topics in reading instruction. This situation is natural enough, for beyond the decoding stage, advancement in reading depends so much on the child's developing language abilities that interact with almost all other instruction and experience. We do, of course, often try to teach specific skills, such as locating the main idea or understanding poetry. We are relatively uncertain about how to teach such skills; they often seem to develop without specific instruction, and they are highly intercorrelated.

A second reason that we are relatively unconcerned about grade changes in the content of reading tests is that the same children who learn the decoding skills readily also typically continue to score well on later tests of richness of vocabulary or inference. There is considerable evidence of this stability of performance. For example, unpublished studies by Joseph Breen show correlations generally in the 70s between reading achievement at the end of grade one or grade two and reading achievement in the fourth or fifth grade (*3*). Now such stability could be taken as evidence that the tasks posed by reading tests really do not change very much from first to fifth grade. I have offered the

high correlation between intermediate grade reading vocabulary tests and verbal aptitude tests as evidence that they are testing about the same thing. The difference in the two cases is partly the evidence of one's eyes. The reading vocabulary sections of a reading test and of a paper and pencil verbal aptitude test look alike. They were prepared following similar principles to test, in printed form, richness of vocabulary. On the other hand, reading vocabulary and comprehension items for the early grades are built on different principles from those for later grades and the result is readily apparent in the items.

There are other considerations to make one reject the high correlation between first and fifth grade reading scores as evidence that items designed to test decoding skills are actually testing the same ability as later items. One of those considerations is that some of the variance in scores at first and second grade level is based on items like those for higher grades. The harder items on second grade tests, at least, are often constructed like those for higher grades, since the norms on such tests extend into the intermediate grade level. Again, this situation results partly from the fact that reading achievement for many children in the intermediate grades is not so dependent on specific school instruction as achievement in some other subjects.

There is another consideration that argues against accepting the high correlation between beginning reading achievement and later reading achievement as evidence that the beginning and later items are measuring the same reading skills. This consideration is that first and second grade reading achievement scores correlate remarkably highly with all kinds of later academic achievement, including arithmetic, not just with later reading achievement. In Breen's studies, correlations between first or second grade reading scores and fourth or fifth grade arithmetic scores were also in the 70s, though somewhat lower than correlations with fourth and fifth grade reading scores. Correlations between first or second grade reading scores and composite scores on the Iowa Test of Basic Skills in fourth or fifth grade were in the 80s. The first or second grade reading items are clearly not arithmetic items. They are simply measuring something that is strongly related to later achievement.

Why are early reading scores so highly related to later school achievement? Do teachers continue to favor children who are ini-

tially favored by them? Do scores on early reading tests influence teachers' expectations and lead to self-fulfilling prophesies? Do homes that provide support for early success in reading continue to provide good support and encouragement for other school achievement? Do children who have the capacity to learn to read easily also have good capacity for other learning? Do children who are adaptable and malleable enough in the school environment to participate well in beginning reading instruction also participate well and thus learn more from later instruction? Does the reading skill itself, and the knowledge gained through using it, contribute so much to school achievement in other subjects that growth in achievement is essentially determined by it? Probably all these things are true in varying degrees. You can undoubtedly add other reasons to the list. My own belief is that, of the possibilities mentioned, perhaps the most important is the continuing and reasonably consistent influence of the home environment. There are great variations in the degree to which the home provides a source of motivation and support, establishes habits of attention and cooperation, provides a background of useful skills and information, and, probably not least in importance, supplies actual instruction on school subjects.

In any case, for whatever reasons, reading ability at the end of first or second grade is highly related to later achievement in reading and other subjects. Put another way, a child who has not learned to read by the end of the second grade is in deep trouble in most school systems; the child who does learn to read in first or second grade finds that he has been planted in a child's garden of reverses. There are exceptions, of course, but most such children are in for a long career of frustration and failure. That there is a strong correlation between early success in reading and later school achievement does not necessarily mean that preventing early reading failures would drastically reduce later school failures. The effects of a prevention program would depend on the reasons for the strong relationship between early reading achievement and later school achievement. On the other hand, we do know that if nothing is done, those children who now do not learn to read in the first two years are very likely to be saddled with failure for the rest of their school careers. It is surely worth a try—worth an all-out effort to see that every child who doesn't make good progress in early reading has every incentive and every

opportunity to learn the skill. I am not suggesting that all children can achieve equally well, simply that the school should recognize what an extremely serious matter it is when a child doesn't learn to read in the first grade or two and that the school should do all that possibly can be done at that time rather than waiting until later.

So far, I have been illustrating the point that the nature of reading achievement tests changes markedly from the first grade to the intermediate grades. Now let us look at the other side of the statement that introduced this point, namely that at a given educational level there is not much difference between reading subtests with different names. Correlations between the vocabulary subtest and the comprehension subtest generally approach the reliability of the individual subtest. There is still room for the two subtests to be measuring somewhat different achievements, but for individual pupils the difference between the vocabulary score and the comprehension score must generally be very large before we can put much faith in this difference actually reflecting a true difference in achievement in the two areas. The same statement applies with even greater force to attempted subdivisions of the vocabulary and comprehension tests. At the intermediate grade level and above, repeated studies of different types of formats of vocabulary testing emphasize that more or less the same achievement is being measured by the different types of vocabulary tests. There is, indeed, some difference, but the value of separate subtest scores for different types of vocabulary tests at the intermediate grade level and above seems questionable at this time.

At the stage of beginning reading, however, there is probably room for more differentiation of the skills that are tested than has so far been incorporated into most tests. Any achievement test should, of course, be directly relevant to what is being taught in the school. At the present time, there is a considerable variation in the way beginning reading is taught. Most reading vocabulary tests for the first two grades include a mix of items for measuring the outcomes of these different emphases. It is probably at these earliest stages of reading instruction that criterion-referenced measurement can be most meaningful and helpful at the present time in assessing reading achievement. At advanced stages of achievement, criteria will be much harder to specify, and if we follow our intuitions in setting them, we are likely to obscure rather than clarify

the problem of the taxonomy of reading ability. Some criteria that will seem to make common sense will not help us understand what skills we need to teach. We need to continue to study this problem of the skills and abilities that comprise reading achievement.

At the intermediate and higher levels, separation of different types of comprehension is about as difficult as separating different aspects of vocabulary achievement. The work of Davis (*1, 2*) clarifies the nature of this problem and indicates some of the potentials that exist. At the present time, the most promising distinction exclusive of vocabulary would appear to be that between understanding facts explicitly stated in the reading passage, and making inferences from what is stated. Even this distinction is not an easy one, and we should require a clear demonstration (such as Davis has been attempting to provide), that two subtests are measuring this distinction before we pay much attention to comprehension subtest scores that claim to represent different aspects of comprehension ability.

Let me now illustrate the significance of the changes in the nature of reading tests from first grade to the later grades by giving an example of how these changes might influence our understanding of test results. It was noted earlier that the reading vocabulary subtest ended up in the intermediate grades being essentially like the vocabulary section of a group intelligence test. Some school systems have recently abandoned the use of so-called intelligence tests on the grounds that they lead to discrimination against pupils whose backgrounds have not equipped them well for traditional school studies. When one evaluates the justification for this step, it becomes evident that the potential harm from the intelligence test lay in its title and in the surplus meaning given to the scores, not in the information it actually provided. It provided information about the student's current ability to learn academic subjects through reading, or listening, to expositions of academic material in standard English. The reading vocabulary test provides that kind of information, too. In fact, a reliable reading test is likely to predict later school achievement about as accurately as an IQ test. But look at the difference in attitudes toward these two test scores. A reading-vocabulary test is looked on as a measure of the school's accomplishment or the school's failure, whereas the vocabulary section of an intelligence test yields a score that is someone else's responsibility. One way of indexing the difference

in attitude toward the two types of tests is to note the difference in the temptation to coach students on the answers to the two. Coaching, and other fraudulent ways of making sure that the reading test scores of a class or school look good, has become a serious problem in some schools. Coaching is ordinarily not a problem for IQ tests given by the school. A low IQ score is taken as an indication that the child will have difficulty in learning. It can even serve as an excuse.

The teacher may not realize that the reading vocabulary test is very like a section of the IQ test. But the teacher does know that a child who scores low on the reading test will have difficulty in learning at school, just as she knows that the child who scores low on an intelligence test is likely to have difficulty learning in school. The teacher will probably assume, however, that the difficulties have different sources and different remedies. She believes that the remedy for the low reading test score and for the difficulties that it indexes is to teach the child to read. She is likely to see a low score on an IQ test as meaning that she can't teach the child to read.

My purpose in raising these questions about the similarity between reading vocabulary and IQ vocabulary tests and about the difference in reaction to them is not to get the reading tests abandoned too. The reading part of a reading test is the comprehension subtest, and surely we do want to know how well children are learning to read. Rather, I wish to point out that similar experiences and similar background factors influence the scores on the reading vocabulary test and on the vocabulary section of the IQ test.

In the past, we have tended to think of the intelligence test score as reflecting the child's past and as indicating the extent to which the school will have trouble teaching him in the future. We have thought of the reading test score as reflecting the school's work in the past and as indicating the extent to which the child will have trouble in the future. We will face more intelligently the tasks of teaching reading and will face with even greater determination the whole job of education when we understand the functions and problems of measurement well enough to realize that both scores reflect the child's past and what the school has done, and that both scores suggest future needs and opportunities for both the child and the school.

References

1. Davis, F. B. "Research in Comprehension in Reading," *Reading Research Quarterly,* 4 (1968), 499-545.
2. Davis, F. B. "Psychometric Research on Comprehension in Reading," *Reading Research Quarterly*, 7 (1972), 628-678.
3. Thorndike, R. L. "Reading as Reasoning," *Reading Research Quarterly,* in press.

Ronald P. Carver
American Institutes
for Research

READING AS REASONING: IMPLICATIONS FOR MEASUREMENT*

In 1917, Edward L. Thorndike (*15, 16*) presented the argument that reading was basically a reasoning process, and in 1971 Robert L. Thorndike (*17*) presented factor analytic and correlational data which he interpreted as supporting this idea. Furthermore, R.L. Thorndike argued that if we desire better readers, the challenge is to develop ways of teaching people to think rather than primarily concentrating on reading. He concluded that it is primarily meager intellectual processes that are limiting reading comprehension, not deficits in one or more specific and readily teachable skills.

Before embarking upon another innovative teaching program designed to improve reading skills by teaching reasoning skills, it seems prudent to take a close critical look at the measurement techniques used to collect the data supporting the reading as reasoning argument. It appears that the close relationship found between reading and reasoning may be an artifact of the measures employed. The primary purpose of this article will be to analyze critically the relationship between reading and reasoning with the aim of illuminating the test and measurement problems involved.

First, a background for discussion of the research by the two Thorndikes will be presented, and then the implications for present day reading tests will be discussed. Finally, suggestions for developing future reading tests will be presented.

*The preparation of this paper was supported in part by the Personnel and Training Programs of the Office of Naval Research, Contract No. N000014-72-C-0240.

BACKGROUND

Before delving into the critical analysis of the Thorndike research, there needs to be established a frame of reference for what is meant by reading. An understanding of the reading process is crucial to any interpretation of the idea of reading as reasoning. An elaborate explication of the reading process seems to be justified. The following quotation from Spache (*12*) helps to convey the extensiveness of what is intended by references to the process of reading comprehension.

> The reader first recognizes words by their form, shape, structural parts or by the implications of the context. Each word calls forth one or several meaning associations which the reader tries out for appropriateness in this contextual setting. He accepts what seems to be the most relevant meaning or associative thought and proceeds to the next word, again choosing an association which seems logically related to the preceding word. Various groups of words form cohesive associations as he reads through the elements of the sentence. These groups of ideas or details coalesce into the stated or implied meaning of the sentence. The meanings of successive sentences may be combined inductively into the main idea of the paragraph. In deriving the main ideas of the paragraph, the reader may recognize cause-effect, question-answer, hypothesis-proof or other relationships which contribute to the generalization. Or these sentence meanings may form the basis of original deductions, such as implications or unstated conclusions, or ideas associated with but tangential to the main idea of the paragraph.
>
> The reader may go far beyond simple comprehension of the literal, implied or tangential meanings to evaluation of the ideas offered. He may question their authenticity, deny their implications, or reject the bias or prejudice present. He may be moved to consult other sources for verification, to check the author's background, to compare the author's value judgments. Finally, the reader may utilize the author's ideas or viewpoint in a creative treatment of the same topic basing his own ideas upon those he has read, or refuting them by proper logic or proof.

It seems convenient to ferret out four separate levels of Spache's total description of the reading comprehension process. Level 1 is associated with the words as units and involves both the decoding of words and the determination of their meaning as used in the particular sentence being read. Level 2 is associated with sentences as units and involves the combination of the meanings of the individual words into the complete understanding of the sen-

tence. Level 3 is associated with the paragraph as a unit and may involve the recognition of the implied main idea of the paragraph. Level 4 is associated with no particular unit and may involve thinking activities which are not at all associated with the literal, implied, or tangential meanings of the prose. It is important to note that Level 4 would seem to involve a great deal of what is normally regarded as reasoning. In fact, Level 4 would not seem to be a part of the ongoing process of reading at all. It seems to involve activities that probably could not occur at the same time as Levels 1 and 2 were occurring. Level 4 is probably best regarded as not being part of the *reading process* at all, although Spache and others may want to include it as part of what they regard as the total *reading comprehension process.* Level 3 also seems to involve a great deal of reasoning. The recognition of main ideas and cause-effect, question-answer, hypothesis-proof relationships would seem primarily to involve basic intellectual processes that may or may not be functioning at the same time as Levels 1 and 2. In any event, because of the inherent nature of the activities that take place during Level 3, it could be assumed that a primary intellectual functioning, called reasoning, would have to be involved. Levels 1 and 2 seem to capture the essence of the ongoing reading process. And it is not at all easy to infer the extent to which a basic intellectual process such as reasoning is involved in the execution of Levels 1 and 2.

It appears that the functioning of each of the levels noted above would depend upon the functioning of every lower level. And the higher the level, the more obvious is the functioning of reasoning. If reading is taken to include Levels 1, 2, 3, and 4, then it is easy to understand how reading could be regarded as involving a great deal of reasoning because of the way Levels 3 and 4 have been defined. What is crucial is the relationship between reasoning and what happens in Levels 1 and 2, the essence of the reading process. When the reading process breaks down, then it becomes crucial to know why. Was it because the level of reasoning ability was not high enough to match the level required by the material? Or was it because the individual had simply not yet learned how to recognize the words and determine their meaning within the sentence? If we understand the relationship between reasoning and reading (Levels 1 and 2), then we shall be in a much better position to diagnose and prescribe when confronted with a dysfunc-

tioning reader.

With this background, we are now ready to examine the research of the Thorndikes (*15, 16, 17*).

THORNDIKE, 1917

E. L. Thorndike seemed to be interested in Levels 1 and 2 of reading. He seemed to want to learn more about the processes involved in reading while it was occurring, *i.e.*, the ". . . dynamics whereby a series of words whose meaning is known singly produces knowledge of the meaning of a sentence or paragraph." Thorndike speculated in great detail about the thoughts which accompany the words while they are being read by a poor reader as compared to an expert reader. Although Thorndike had a primary interest in Levels 1 and 2 of reading, his research did not seem to concentrate on Levels 1 and 2 as they are executed in normal reading situations. As Tuinman (*18*) has noted, Thorndike was highly interested in showing that reading fits into the prevailing stimulus-response psychology of the time. He was interested in demonstrating that reading could involve the higher intellectual processes such as reasoning just as much as does mathematical calculation. Thus, we find Thorndike using extremely difficult reading material (*13*). It should not go unnoticed that Thorndike, in this research, did not use paragraphs that had been taken from existing school reading materials. He devised his own paragraphs and because of this and because of the nature of the paragraphs, it is questionable whether one should agree that they were representative of the ordinary reading done by his subjects, as he contended (*14*). The paragraphs indeed seem like exercises in logic, and when one examines them it is not difficult to understand how he could conclude that reading involved reasoning.

Besides the passages chosen by Thorndike, there are other reasons why his research does not seem to provide evidence relevant to the relationship between reasoning and Levels 1 and 2 of reading. Thorndike's research task involved passages and questions on the passages. There is a problem in making valid inferences about the ongoing reading process from the answering of certain questions presented subsequent to the reading itself. If an individual does not answer the question correctly, it might be because of a deficit in the reasoning process that occurred during reading (as Thorndike seems to infer), it might be because of a deficit in the

reasoning process that occurred in connection with an attempt to answer the question itself, or it might be because a failure to execute Level 1 of the reading process made it impossible to answer a Level 3 or Level 4 type question. Thorndike attempted to deal with this problem by considering the passage and the questions as a single unit so that he could increase difficulty or degree of understanding either by replacing paragraphs or rewording questions (*18*). Since questions are not an inherent appendage to reading, it seems prudent to conclude that Thorndike was not using a research paradigm which allowed optimum generalization to the reading process.

There is direct evidence that E. L. Thorndike used Level 3 type questions in his research. For example, the very first question he presents in support of his argument is: "What is the general topic of the paragraph?" Another example of the type of questions he asked pupils in Grade 6 is: "What condition in a pupil would justify his nonattendance?"

It also appears that Thorndike did not adequately control for failures in Level 1 reading, the part that would seem least to involve reasoning. These same failures could also involve what is known as decoding problems today. Thorndike (*16*) was aware of this problem as the following quotation demonstrates:

> In general, the material used here will be paragraph and questions whose words singly are fairly well known to the pupils in question, but whose sentence structure is somewhat more elaborate than pupils of the grade in question can manage. That is, the study is primarily concerned with the ability of the pupil to understand totals, few of whose elements are unknown, but whose internal relations are somewhat intricate and subtle.

This plan seems reasonable, but consider the following words Thorndike used in the paragraph and questions he administered to sixth graders in 1917: session, contagious, impassable, compulsory, and excusable. There is no direct evidence that all or even most of these sixth graders knew these words (i.e., could execute Level 1).

In summary, Thorndike seemed to be interested in ordinary reading (i.e., the functioning of Levels 1 and 2), but his research seemed to: 1) involve unordinary reading materials; 2) include questions that were definitely reasoning type questions (i.e., Level

3 reading type questions); and 3) inadequately control for Level 1 dysfunctions accounting for Levels 2, 3, and 4 failures. Because of these aspects of his research, it seems reasonable to conclude that Thorndike's research contributed little or nothing to our knowledge of the relationship between reasoning and the primary aspects of the reading process, i.e., Levels 1 and 2. It seems easy to agree that reading involves high degrees of reasoning when reading is taken to include Levels 3 and 4, but this is not a relationship—it is a definition.

THORNDIKE, 1971

R. L. Thorndike presented a paper entitled "Reading as Reasoning," upon receipt of the Edward L. Thorndike Award at the 1971 meeting of the American Psychological Association (*17*). Included in the various data analyses of R. L. Thorndike was a factor analysis of the data presented by Davis (*7*). Thorndike concluded that one factor, reasoning, could be interpreted as accounting for the predominant portion of the variance. This analysis and conclusion is in agreement with that of Carver (*3*), who reported that the intercorrelations among the Davis variables were above .90 when the reliability coefficients were used to correct for attenuation.

R. L. Thorndike has presented a great deal of data which convincingly shows that the ability to answer questions on existing standardized reading tests is so highly correlated with achievement tests and intelligence tests that it is reasonable to conclude that they are measuring the same thing. Yet, the high correlations among reading and intelligence tests seem best to be regarded as artifactual and as having little or nothing to do with the nature of reading. The reasons for this involve the way the tests are made.

The intelligence tests in Thorndike's research were group tests and all, therefore, required reading. Thus, these measures of basic intellectual functioning were all contaminated to an unknown but presumably high degree by variations in reading ability.

More importantly, the standardized reading tests used by R. L. Thorndike were all contaminated to an unknown but presumably high degree, by intelligence (reasoning-type) questions. Farr (*9*) has recently discussed and given examples of how reading tests today bear a strong resemblance to group verbal intelligence tests.

One of Farr's examples taken from a well-known reading achievement test is given below:

The sheep were playing in the woods and eating grass. The wolf came to the woods.

Then the sheep

1. went on eating.
2. ran to the barn.
3. ran to the wolf.

The reading tests of today make no effort to discriminate between questions relevant to Levels 3 and 4 and questions relevant to Levels 1 and 2 when they select and score test items for passages. Since traditional item selection techniques involve the selection of the items which best discriminate among individuals, and since intelligence or reasoning-type items tend to be the best in this regard, it is not surprising that standardized reading tests have evolved into standardized verbal intelligence tests. It is unfortunate but true that E. L. Thorndike's research has provided the justification for permitting reading tests to become reasoning tests. If questions that are clearly reasoning-type questions make up much, if not most of present day standardized reading tests, then it should not be surprising to find that reading test scores are reasoning scores. Yet, we shall not learn much about the reading process, Levels 1 and 2, by employing a task that reflects directly upon Level 3 and Level 4 reading, which are already known to involve high degrees of reasoning.

Up to this point it has been argued that it is not surprising that if reading is measured by passages and questions that obviously require reasoning, then reading is bound to appear to be reasoning. Yet, informally, colleagues have rebutted that even questions that do not seem to be reasoning-type questions correlate highly with those that are. This purported inconsistency also may be an artifact of the way tests are developed. For example, consider the Davis (7) study which had a variable called "finding answers to questions answered explicitly or merely in paraphrase in the content." This variable would appear to be an indicant of reading, Level 2. And this variable seems to be just as much a reasoning-type variable as the others which were more obviously reasoning-type variables. Yet, there is a major problem involved in the interpretation of this inconsistent result. Not only does one have to make sure the question is not of a reasoning-type, but one also has

to make sure that the alternative wrong answers provided on a multiple-choice test do not inadvertently shunt the item off into a reasoning-type question.

The above mentioned problem is a serious one, given existing item writing and selecting techniques. If an alternative wrong answer does not draw any responses (i.e., it is a poor distractor), it is usually rewritten so that it becomes more credible, i.e., more people choose it. Thus, a test question may appear to require little or no reasoning, but when "good" alternative wrong answers are provided, then the test may be automatically changed so that it requires varying degrees of reasoning not obvious from the question itself. The degree to which this is true in the Davis research is not known. Yet, this is an inherent problem involved in all multiple-choice tests, and research which was not designed to control for this artifact should be interpreted with caution in regard to the relationship between reading and reasoning.

R. L. Thorndike concluded from his results that reading was fundamentally reasoning, and he further suggested that improvement in reading may only occur after instruction in reasoning. Yet, it seems more reasonable to interpret R. L. Thorndike's results as supporting an alternative hypothesis. Reading is not primarily reasoning, but most standardized ***reading*** tests are actually standardized ***reasoning*** tests. The high correlation between reading tests and intelligence tests reported by E. L. Thorndike is still true today. For example, most of the correlations between the STEP Reading Test and the SCAT Test (an intelligence test) are reported to be above .80, according to the manual for the test. But, it seems more reasonable to interpret these high correlations as artifacts of the way the tests are developed rather than supporting the idea that reading is primarily reasoning.

A CRITIQUE OF STANDARDIZED READING TESTS

E. L. Thorndike's technique of presenting paragraphs with questions beside them has influenced standardized testing of reading achievement to the present day, as evidenced by the tests used in R. L. Thorndike's research. What has changed through the past fifty years is the addition of multiple-choice answers and highly sophisticated ways of revising, scoring, analyzing, and reporting test results. It does not matter to most psychometricians how

good the test questions are for measuring progress in the understanding of the sentences that occurred during the reading of a paragraph; i.e., as long as the questions have face validity, discriminate reliability among individuals at any given level, and demonstrate level-to-level group mean increments. It was fortunate for psychometricians that E. L.Thorndike focused upon the reasoning aspects of reading, since it is quite easy to develop tests that satisfy the preceding psychometric criteria using reasoning-type questions. For school-age individuals, intelligence-type questions always produce large individual differences and show maturational increases from year to year. If E. L.Thorndike had focused upon questions that did not produce large individual differences, psychometricians would have had large problems adapting their sophisticated statistical techniques to the development of reading tests.

To illuminate this undesirable influence of E. L. Thorndike upon present day measures of progress in reading, a hypothetical situation will be presented. Suppose a paragraph is selected from a sixth grade reading book, and questions are constructued which are designed to ascertain whether a student has read and understood the complete thoughts (i.e., sentences) that the writer intended to communicate (reading, Level 1 and Level 2). Suppose most of the sixth graders can get all of these questions correct; i.e., the variability in the group approaches zero. In this situation, all traditional estimators of reliability and validity will also approach zero (*4*). The traditional psychometrician will throw up his hands in horror in this situation. Yet, this type of test situation may be the best way to measure progress in reading. (Empirical data have been presented to support this measurement method; see *6*.) If a test on a paragraph measures levels of progress, then the variability among individuals may not be found primarily on the test but in the time or amount of instruction required for the individual to reach this level of mastery (*2*). Yet, variability in time is anathema to the traditional calculation of test percentiles and reliability estimates.

Unfortunately, E. L.Thorndike and R. L.Thorndike continue to influence psychometricians to be unconcerned about the representativeness of their questions for indicating progress in reading (Level 1 and Level 2). Consider, for example, the following categories of items given in the manual for the STEP Reading Test:

reproduce ideas, translate ideas and make inferences, analyze motivation, analyze presentation, and criticize. The understanding that occurs as a result of the execution of Level 1 and Level 2 may be more difficult to measure than the reasoning that occurs during Level 3 and Level 4, but that does not seem to be justification for allowing reading tests primarily to measure important but ancillary aspects of the reading process. It would seem to be a better strategy to let intelligence tests measure reasoning and reading tests measure reading, Level 1 and Level 2. Why be inefficient and duplicate these measures even if some reading specialists do consider reading comprehension to include the verbal reasoning involved in Levels 3 and 4? At present, we are in the embarrassing position of assigning certain grade level of reading achievement (e.g., Grade Level 3.4) to a chance score on a test and not really knowing if the student can read (Levels 1 and 2) at all. Or, the chance score may indicate that the questions (Levels 3 and 4) were so irrelevant to the reading process that the student understood the material that he read but could not sufficiently infer what the answers to the questions were.

It is haunting to consider that today's standardized reading tests are probably measuring *reasoning* progress more than *reading* progress. Consider a pair of hypothetical twins: Twin A receives no instruction in reading for an entire year, and Twin B receives normal reading instruction. Twin A no doubt matured one year in reasoning ability whether he received instruction in reading or not. Twin A probably will show as much, or almost as much, progress on most standardized reading tests as Twin B. Thus, a school which has a poor instructional program will probably demonstrate about as much gain in a year's time as a school which has an excellent program, a haunting thought!

It is especially frightening to find that the U.S. Office of Education contracted with Educational Testing Service for a National Anchor Test Equating Study in Reading (*10*). This contract was for $698,000.* The results of this study most likely will be a

*This sum of money is approximately equal to the total amount of funds to be expended during the entire 1972 fiscal year in the Basic Research Program of the U.S. Office of Education. Reading researchers should be interested in the fact that USOE gave psychometricians $698,000 to make the centiles of seven norm-referenced tests more comparable, while USOE earmarked no federal funds for basic research in reading. USOE did earmark funds for basic research in economics and anthropology, but the USOE initiated Targeted Research and Development Program in Reading received no USOE funds.

highly accurate, norm-referenced intelligence test system under the guise of a reading test. After this study, it will seem logical that school systems will be forced by USOE to evaluate their federally funded innovative reading programs with one of the seven most popular tests that have been nationally equated. We have already witnessed the deprecation of preschool programs because they did not raise IQ scores; even though no one should expect IQ scores to be raised (*11*). Should we now expect to see innovative reading programs bite the dust when they are eventually subjected to psychometrically sound tests that primarily measure progress in reasoning? It seems rationally sound to expect school systems to help students in their learning to read more difficult material, but it seems rationally unsound to expect good reading instruction to have much effect upon a student's fundamental ability to reason.

THE FUTURE OF STANDARDIZED READING TESTS

It is known that the ability of almost every student to think or reason (e.g., mental age) increases each year throughout school age, and it is known that reading skill increases normally for some and does not increase normally for others. It is not known whether certain levels of basic intellectual ability are required for certain levels of reading achievement. It may or may not be realistic to expect that almost all 10-year-olds can achieve a certain minimum level of adult reading skill given ample time and sufficient help (*8*).

Bloom (*1*) contended that there are certain hurdles in school that should be overcome before an individual is subjected to subsequent higher level instructional treatment. Otherwise, the student may never progress normally. One of the challenges in reading is to measure the achievement of these hurdles. Another challenge is to determine how much time an individual needs to achieve a hurdle, given a certain developmental level of reasoning. Just because a student is low in reasoning ability in relation to his same-age peers does not mean that his reasoning ability will not improve as he grows older, or that he should not be expected to attain a certain level of reading ability to match each level of his reasoning ability as it matures.

What appears to be needed are: 1) tests that actually measure

progress levels in the ability to read, i.e., edumetric* or criterion-referenced type tests that focus upon the ability to read and understand reading material of increasing difficulty instead of psychometric or norm-referenced reasoning-type reading tests; and 2) tests of edumetric or criterion-referenced levels of reasoning ability instead of psychometric norm-referenced reasoning tests. Then, answers to the following theoretical and practical questions about the relationship between reading and reasoning could be empirically determined.

1. Does the rate of growth in reading match the rate of growth in reasoning?
2. Does the level of reasoning ability always determine the highest level of reading ability?
3. Can the level of reasoning ability be used to set the expectation level of reading status and thereby be used to evaluate the progress of the students and the "goodness" of the school system's instructional program?

E. L. Thorndike in 1917, and R. L. Thorndike in 1971, have focused upon an important problem, but the past, present, and future ill-effects of this focus should not be overlooked. What is needed at this time is more attention directed toward the measurement of absolute levels of the ability to read sentences that make up paragraphs, not the ability to answer reasoning-type questions on paragraphs. What is needed is an investigation of the relationship between absolute levels of reading and absolute levels of reasoning. Hopefully, the next fifty years will not find reading researchers in the same embarrassing situation of concluding from reading test data that the ability to answer reasoning-type questions on paragraphs mainly involves the ability to reason.

*The edumetric approach to testing refers to the focus upon measuring progressive within-individual gains of high relevance to education as contrasted with the traditional psychometric approach which tends to focus upon the static between-individual differences of high relevance to psychology (5).

References

1. Bloom, Benjamin. "Individual Differences in School Achievement: A Vanishing Point?" Phi Delta Kappa Address at the meeting of the American Educational Research Association, New York, 1971.
2. Carroll, John B. "A Model of School Learning," ***Teachers College Record,*** 64 (May 1963), 723-733.
3. Carver, Ronald P. "Analysis of 'Chunked' Test Items as Measures of Reading and Listening Comprehension," ***Journal of Educational Measurement,*** 7 (Fall 1970), 141-150.
4. Carver, Ronald P. "Special Problems in Measuring Change with Psychometric Devices," Proceedings of the A.I.R. Seminar on Evaluative Research, Strategies and Methods. Pittsburgh: American Institutes for Research, 1970.
5. Carver, Ronald P. "Reading Tests in 1970 versus 1980: Psychometric versus Edumetric," ***Reading Teacher,*** 26 (December 1972), 299-302.
6. Carver, Ronald P. "Measuring the Relationship between Reading Input and Understanding," unpublished manuscript, 1973.

   Davis, Frederick B. "Research in Comprehension in Reading," ***Reading Research Quarterly,*** 3 (Summer 1968), 499-545.
8. Ellson, Douglas G. "A Critique of the Targeted Research and Development Program on Reading," ***Reading Research Quarterly,*** 5 (Summer 1970), 524-533.
9. Farr, Roger. "Measuring Reading Comprehension: An Historical Perspective," in F.P. Green (Ed.), ***Twentieth Yearbook of the National Reading Conference.*** Milwaukee: National Reading Conference, 1971, 187-197.
10. Jaeger, Richard M. "A National Test Equating Study in Reading," paper presented at the meeting of the Psychometric Society, St. Louis, April 1971.
11. Jenson, Arthur R. "How Much Can We Boost IQ and Scholastic Achievement?" ***Harvard Educational Review,*** 39 (Winter 1969), 1-123.
12. Spache, George D. ***Toward Better Reading.*** Champaign, Illinois: Garrard, 1963, 65.
13. Stauffer, Russell G. "Thorndike's 'Reading as Reasoning': A Perspective," ***Reading Research Quarterly,*** 6 (Summer 1971), 443-448.
14. Thorndike, Edward L. "An Improved Scale for Measuring Ability in Reading," ***Teachers College Record,*** 16 (November 1915), 31-53.
15. Thorndike, Edward L. "Reading as Reasoning: A Study of Mistakes in Paragraph Reading," ***Journal of Educational Psychology,*** 8 (June 1917), 323-332.
16. Thorndike, Edward L. "The Understanding of Sentences: A Study of Errors in Reading," ***Elementary School Journal,*** 8 (October 1917), 98-114.
17. Thorndike, Robert L. "Reading as Reasoning," ***Reading Research Quarterly,*** in press.
18. Tuinman, J. Jaap. "Thorndike Revisited—Some Facts," ***Reading Research Quarterly,*** 7 (Fall 1971), 195-202.

Robert L. Thorndike
Teachers College
Columbia University

DILEMMAS IN DIAGNOSIS

One of the common uses of psychometric devices in the field of reading—as in education generally—is for educational diagnosis. Diagnosis is most often a matter that relates to a specific individual, though we may be from time to time interested in making diagnostic judgments about groups. Diagnostic judgments are often based on the comparison of two measures in order to judge whether the individual shows some genuine discrepancy in the traits or characteristics that the two measures represent. Thus, if a child falls at the 50th percentile on a test of word knowledge but only at the 25th percentile on a test of comprehension of connected prose, the diagnostician must decide how much confidence to place in the conclusion that this child's ability to read connected prose falls short of his knowledge of word meanings. The whole armamentarium of diagnostic devices in the field of reading has its value in suggesting judgments of the type "Ability A is greater than Ability B."

But differential judgments about individuals are slippery customers. They are peculiarly subject to measurement error. Some 45 years ago, Kelley (*2*) warned of the need for especially reliable tests if such diagnostic judgments were to be made with confidence. Nothing that has developed since then has given occasion for the psychometrician to change his views on this point. The diagnostician, however, cannot wait for the psychometrician to produce the perfect psychometric instrument in order to deal with the practical problems of his day to day functioning. He must get on with the job. And practical limitations of time and resources

for carrying out his assessments mean that he will always have to use tools that fall short of psychometric ideals.

This being so, what help can psychometrics offer the diagnostician to "carry on" while he waits for the perfect diagnostic battery? Perhaps some guidance on the level of confidence that he should place in diagnostic judgments might be useful to tide him over.

We must always remember that any test, or any other type of behavior-observation, represents only a limited sample from some domain of behavior. It represents the domain only imperfectly, and the score that it produces is only an approximation to the score that the individual would get for the whole domain—or more realistically, that he would get on other samples drawn from that domain. We get evidence on this variability from sample to sample of behavior through the various procedures for obtaining a reliability coefficient, and we express it most usefully for our present purposes as a standard error of measurement. The standard error of measurement may be thought of as the standard deviation of a series of equivalent measures of the same individual, displaying the extent to which the measures scatter away from his "true score."

Suppose, now, we have two measures, X and Y. For concreteness let us say that X is a measure of word knowledge and Y a measure of paragraph comprehension. Suppose that results from the two measures are expressed in a common equal-unit score scale, such as T-scores or stanines for a common sample of sixth grade pupils. Suppose that Peter differs on the two tests by an amount D, and for concreteness let us say that this difference is 10 points on the T-score scale or 2 points on the stanine scale; i.e., a difference of exactly one standard deviation. How much confidence should we have that this difference represents something real, and didn't just happen because of errors of measurement in the two tests? How confidently can we expect a difference in the same direction, though obviously not of identically the same amount, if Peter is retested with equivalent forms of each of the two tests?

In setting our level of confidence, we need to take account of three things, two of which have already been mentioned. In the first place, we need to take account of the size of the standard errors of measurement for the two variables. The larger the errors—that is, the lower the reliability—the lower the confidence.

The appropriate degree of confidence depends secondly upon the size of the observed difference between the two scores. The larger the difference, the greater the level of confidence. It depends finally, and quite critically, on the correlation between the measures, X and Y, of the two attributes that we are studying. The higher that correlation, the other two factors remaining the same, the less confidence one can have in the meaningfulness of the difference.

Let us look at the rationale for these relationships with specific figures for a definite example. Suppose that the word knowledge test (X) and the paragraph reading test (Y) are each known to have reliability coefficients of .90 for a sixth grade sample and that for the same sample the correlation between the two tests is 0.80. Consider Peter, who scored one standard deviation lower (relative to the standardization group) on the paragraph test than on the word test.

For a single test with reliability of 0.90, the standard error of measurement, expressed in standard deviation units, is:

$$\sqrt{1 - r_{11}} = \sqrt{1 - 0.90} \qquad (1)$$

For the difference between two tests, both expressed in standard deviation units, the standard deviation of differences arising purely from measurement errors, which we might call the standard error of measurement of difference, is:

$$\sqrt{2 - r_{xx'} - r_{yy'}} = \sqrt{2 - .90 - .90} = \sqrt{0.20} = 0.45. \qquad (2)$$

Thus, a difference between scores of one standard deviation is equal to

$$\frac{1.00}{0.45} = 2.22$$

standard errors of measurement of the difference. Turning to tables of the normal curve, we find that a difference this large or larger could be expected to occur in 13 cases out of 1000.

A parallel formula gives the standard deviation of differences between two tests when one knows what the correlation between the two tests is. When, as before, each test's scores are expressed in standard deviation units, the formula for standard deviation of differences is:

$$\sqrt{2 - 2r_{xy}} = \sqrt{2 - 2(0.80)} = \sqrt{.40} = 0.63. \qquad (3)$$

Thus, a difference between scores of one standard deviation is equal to

$$\frac{1.00}{0.63} = 1.59$$

standard deviations of the differences between these two quite highly correlated variables. Turning once again to our table of the normal curve, we find that, given a correlation of this size, differences this large will occur in 56 of 1000 cases. Of these 56, on the basis of our earlier calculation, we should expect that 13 were the result of nothing more than measurement error. This leaves 43 that represent presumably "real" differences. Thus, we may say that the odds are 43 to 13 or about 3 to 1 that the difference is a genuine one. The betting odds of 3 to 1 represent one way of expressing the confidence that we should feel in the diagnostic judgment that Peter is better at word knowledge than at paragraph reading.

Following the same rationale that we have used in our illustration, it is possible to prepare tables showing the "betting odds" for representative combinations of reliability, intercorrelation, and size of difference. An illustrative set of such tables is presented in Table 1.

TABLE 1.

Confidence Tables for Diagnostic Judgments: Odds that an Observed Difference between Two Variables Is a Real Difference.

Section I: Average Reliability = 0.98

Difference in S D Units	Correlation Between Variables									
	.95	.90	.85	.80	.75	.70	.60	.50	.40	.00
0.25	1:1	2:1	2:1	5:2	5:2	5:2	3:1	3:1	3:1	3:1
0.50	9:1	20:1								
0.75										
1.00										
1.25										
1.50	All others greater than 20 to 1									
1.75										
2.00										

Section II: Average Reliability = 0.95

Difference in S D Units	Correlation Between Variables									
	.90	.85	.80	.75	.70	.65	.60	.50	.40	.00
0.25	1:3	1:2	3:5	2:3	3:4	3:4	4:5	5:6	7:8	1:1
0.50	5:4	2:1	5:2	3:1	7:2	7:2	4:1	4:1	9:2	5:1
0.75	7:2	8:1	11:1	14:1	16:1	18:1	20:1			
1.00	13:1									
1.25										
1.50										
1.75					All others greater than 20 to 1					
2.00										

Section III: Average Reliability = 0.90

Difference in S D Units	Correlation Between Variables									
	.85	.80	.75	.70	.65	.60	.55	.50	.40	.00
0.25	1:7	1:5	1:4	2:7	1:3	1:3	2:5	2:5	2:5	1:2
0.50	1:3	1:2	3:4	1:1	1:1	7:6	5:4	4:3	3:2	7:4
0.75	4:5	3:2	2:1	5:2	3:1	3:1	7:2	4:1	4:1	5:1
1.00	3:2	3:1	5:1	7:1	8:1	9:1	10:1	11:1	13:1	17:1
1.25	3:1	7:1	12:1	18:1						
1.50	7:1	20:1								
1.75										
2.00					All others greater than 20 to 1					

Section IV: Average Reliability = 0.85

Difference in S D Units	Correlation Between Variables									
	.80	.75	.70	.65	.60	.55	.50	.45	.40	.00
0.25	1:18	1:9	1:7	1:6	1:5	2:9	2:9	1:4	1:4	1:3
0.50	1:6	1:3	2:5	1:2	1:2	2:3	2:3	3:4	4:5	1:1
0.75	1:3	2:3	1:1	8:7	4:3	3:2	5:3	7:4	13:7	5:2
1.00	2:3	4:3	2:1	5:2	3:1	10:3	11:3	4:1	4:1	6:1
1.25	1:1	5:2	4:1	5:1	6:1	7:1	8:1	9:1	10:1	15:1
1.50	5:3	4:1	8:1	10:1	14:1	17:1	20:1			
1.75	9:2	13:1								
2.00	7:1				All others greater than 20 to 1					

Section V: Average Reliability = 0.80

Difference in S D Units	Correlation Between Variables								
	.75	.70	.65	.60	.55	.50	.45	.40	.00
0.25	1:19	1:11	1:9	1:8	1:7	1:6	1:6	1:5	1:4
0.50	1:8	1:5	1:4	1:3	2:5	2:5	1:2	1:2	2:3
0.75	1:4	1:2	2:3	3:4	4:5	1:1	1:1	9:8	3:2
1.00	2:5	4:5	1:1	7:5	8:5	9:5	2:1	11:5	3:1
1.25	5:8	5:4	2:1	5:2	3:1	7:2	4:1	9:2	7:1
1.50	1:1	2:1	3:1	9:2	11:2	7:1	8:1	9:1	13:1
1.75	3:2	7:2	6:1	8:1	11:1	14:1	17:1	19:1	39:1
2.00	2:1	6:1	11:1	16:1	All greater than 20 to 1				

Section VI: Average Reliability = 0.75

Difference in S D Units	Correlation Between Variables							
	.70	.65	.60	.55	.50	.45	.40	.00
0.25	1:33	1:18	1:13	1:11	1:10	1:9	1:8	1:5
0.50	1:12	1:8	1:5	1:4	2:7	2:7	1:3	1:2
0.75	1:6	2:7	2:5	1:2	3:5	2:3	2:3	1:1
1.00	1:4	1:2	2:3	6:7	1:1	6:5	4:3	2:1
1.25	2:5	3:4	1:1	7:5	5:3	2:1	9:4	4:1
1.50	1:2	1:1	2:1	7:3	3:1	7:2	4:1	7:1
1.75	4:5	8:5	5:2	7:2	9:2	6:1	7:1	14:1
2.00	7:6	5:2	4:1	6:1	8:1	11:1	13:1	30:1

Section VII: Average Reliability = 0.70

Difference in S D Units	Correlation Between Variables						
	.65	.60	.55	.50	.45	.40	.00
0.25	1:47	1:23	1:16	1:14	1:12	1:11	1:6
0.50	1:20	1:10	1:7	1:5	1:5	1:4	2:5
0.75	1:10	1:6	1:4	1:3	2:5	1:2	3:4
1.00	1:6	1:3	1:2	3:5	7:10	4:5	7:5
1.25	1:4	1:2	5:7	1:1	1:1	4:3	5:2
1.50	1:3	5:7	1:1	3:2	9:5	2:1	4:1
1.75	1:2	1:1	8:5	2:1	3:1	7:2	8:1
2.00	2:3	3:2	5:2	7:2	9:2	5:1	14:1

Section VIII: Average Reliability = 0.60

Difference in S D Units	Correlation Between Variables .55	.50	.45	.40	.00
0.25	1:55	1:35	1:24	1:20	1:9
0.50	1:30	1:14	1:10	1:8	1:4
0.75	1:13	1:7	1:5	1:4	1:2
1.00	1:9	1:5	2:7	2:5	4:5
1.25	1:7	2:7	2:5	5:9	13:10
1.50	1:5	2:5	3:5	4:5	2:1
1.75	2:7	3:5	4:5	6:5	3:1
2.00	1:3	3:4	6:5	5:3	5:1

Consider first Section III of Table 1—the section for an average reliability of 0.90—since 0.90 is a fairly representative reliability for good quality ability tests. Note first that no column is shown for an intercorrelation of .90 or higher between the two tests. Whenever the intercorrelation of two tests is as high as their respective reliabilities, they are effective measures of identically the same trait. Differences between the two are then equivalent to (and equal in number to) differences arising solely from measurement error; there is no basis for a diagnostic judgment, and any diagnostic statement should be made with exactly zero confidence.

Note next that when the difference is small, the betting odds are low that this is a real difference, no matter what the correlation. In the row corresponding to a difference of a quarter of a standard deviation, the odds that the difference is a "real" one range from 1 real difference to 7 chance differences when the correlation is 0.85, to 1 real difference to 2 chance differences when the correlation is zero. Most small differences are readily attributed to measurement errors, and our confidence that there is any "real" difference must be correspondingly low.

Finally, in this table we can see the role that the correlation between two test scores plays in our confidence in the reality of any observed difference. This is seen perhaps as clearly as anywhere in the row corresponding to one full standard deviation of difference—a difference that would correspond roughly to falling at the 70th percentile of a group on one measure and the 30th on the other. For a difference of this size, our betting odds would be

3 to 2 in favor of a "real" difference if the correlation between the two test scores were 0.85, 3 to 1 if the correlation were 0.80, 9 to 1 if the correlation were 0.60, and 17 to 1 if the correlation were zero. The confidence we should have in a diagnostic judgment rises sharply as the correlation between the two measures on which the judgement is based decreases.

To view the effect of test reliability on the confidence appropriate for our judgments, it helps to arrange the tables in a somewhat different way. Table 2 shows the "betting odds" when the size of the difference between X and Y is fixed at one standard deviation, but the values of the average reliability and the intercorrelation are allowed to vary. This table makes it emphatically clear how crucially one's confidence depends upon the reliability of the measuring instruments. If the average of the two reliabilities is 0.98 (one should live to see the day when such measures are available!), even the smallest differences, i.e., those of a quarter of a standard deviation, can be accepted with great confidence as real and not the result of measurement error. With a reliability as low as 0.60, a full standard deviation of difference justifies betting odds of less than even money, even when the correlation between the two measures is zero. For intermediate reliabilities, considerable confidence is justified if the correlation between the two measures is low, relatively little confidence is justified if the intercorrelation approaches anywhere near the reliability.

TABLE 2.

Odds that an Observed Difference of One Standard Deviation between Two Variables Is a Real Difference

	Correlation Between Variables										
Average Reliability	.95	.90	.85	.80	.75	.70	.65	.60	.50	.40	.00
.98	All greater than 20 to 1										
.95	13:1	Remainder greater than 20 to 1									
.90		3:2	3:1	5:1	7:1	8:1	9:1	11:1	13:1	17:1	
.85			2:3	4:3	2:1	5:2	3:1	11:3	4:1	6:1	
.80				2:5	4:5	1:1	7:5	9:5	11:5	3:1	
.75					1:4	1:2	2:3	1:1	4:3	2:1	
.70						1:6	1:3	3:5	4:5	7:5	
.60								1:5	2:5	4:5	

What do the tables that we have looked at imply when this type of thinking is carried over to some samples of actual tests with the reliabilities and intercorrelations that characterize them?

Davis (*1*) has carried out some of the most meticulous research on the differentiability of different types of reading skills. Among the abilities that he studied, two that were most readily distinguishable were word knowledge and drawing inferences. His tests had to be quite short, since he was measuring some eight different aspects of reading, so the reliabilities of these two tests were only .58 and .59. The correlation between them had an average value of .45 in several sets of data. Given these values, the betting odds are only 1 to 4 that a difference of one standard deviation between scores on the two tests is "real"; for a difference of two standard deviations the betting odds are 9 to 8. As they stand, the tests hardly justify diagnostic inferences even when the differences are *very* large. But these tests were short—only 12 items each. If they were lengthened to 48 items, which might be a reasonable length for a test in practical use, one estimates that the reliabilities would be increased to .85 and .86, and the intercorrelation to .66. Then the betting odds are respectively 5 to 2 for a difference of one standard deviation and 80 to 1 for a difference of two standard deviations. Thus, we see how very critically diagnostic inferences depend upon the reliabilities of the constituent measures.

Two of Davis' tests that measure more similar functions are the test of inference and a test that calls for identification of the author's tone, mood, and purpose. Here the reliabilities are .59 and .63, and the intercorrelation is .52. Given those values, the betting odds for the existing test are only 2 to 11 for a difference of one standard deviation and 3 to 5 for a difference of two standard deviations. Lengthened to 48 items, reliabilities become .84 and .88 and the intercorrelation 0.75. For this lengthened test, the betting odds are 5 to 3 that an observed difference of one standard deviation is "real" and 23 to 1 for a difference of two standard deviations.

Let us turn our attention now to the Stanford Diagnostic Reading Tests, the distinctive value of which is presumed to lie in their diagnostic effectiveness. Here, unfortunately, the manual provides only single-testing estimates of reliability, and these are certainly somewhat inflated. We cannot know how much. If we take the figures at face value, the average of the subtest reliabilities

is 0.90 and the average of the subtest intercorrelations is 0.65. A more realistic estimate of alternate-form reliabilities might be 0.85. If we assume that figure, and turn to Section IV of Table 1 for reliability 0.85, we find figures in the column for intercorrelations of 0.65 as follows:

0.25 S.D. (which would occur for 38% of children)	1 to 6
0.50 S.D. (which would occur for 27% of children)	1 to 2
0.75 S.D. (which would occur for 19% of children)	8 to 7
1.00 S.D. (which would occur for 12% of children)	5 to 2
1.50 S.D. (which would occur for 4% of children)	10 to 1
2.00 S.D. (which would occur for 1% of children)	over 20 to 1

Thus, if we limit our diagnostic inferences to the one percent with the most extreme differences, our judgments will almost always have a real basis. If we set a lower threshold, and undertake diagnostic statements for as many as 10 percent of children, there will be a basis in reality for something like three-fourths of our judgments. If we set a still more liberal standard, and venture diagnostic statements based on observed differences for as many as 20 percent of the group, the statements will correspond to real differences only about half the time.

Finally, consider a set of data for the reading test of the Stanford Achievement Battery given once in the sixth and once in the eighth grade. For one suburban New York school system, the correlation between the two testings was .747. An estimate of reliability drawn from the test manual is .93. How much would a child have to change his position in his group from the first to the second testing for us to have an even-money bet that there was a real change? The answer comes out to be 0.40 standard deviations. If a child were to improve his position in his group by four-tenths of a standard deviation (for example, from the 50th to the 65th percentile), it is a fifty-fifty proposition that this represents some degree of real change and not just the effect of measurement errors.

The tables and illustrations that we have examined illustrate the impact of reliability, intercorrelation, and score difference upon the confidence that one can logically place in an observed difference between two scores. They illustrate that over the realistic range of test reliabilities, and using the kinds of pairs of measures that we are likely to want to use in diagnostic studies, the confidence is often distressingly low. But children with reading

disabilities are there, and they won't just go away until that happy day when we have diagnostic tools of reliability high enough to permit us to make judgments of score difference at a high level of confidence. Therein lies our dilemma. Wherein do we find our salvation?

If salvation exists, it lies in the fact that most of the actions following from diagnostic judgments are reversible, and if they are unfounded they are likely to result in wasted time or effort rather than any more crucial loss. In this respect, instructional decisions differ from selection and classification decisions, since these are typically permanent. The young person who is denied access to a particular educational institution or job is not likely to be given a second chance. But if the special instruction in word-analysis skills that seems to be called for by a diagnostic reading profile is not effective, it is always possible to hold up, take stock, get new or additional evidence, and follow up some alternative hypothesis. Our tables of betting odds suggest how tentative our hypotheses should often be. Fortunately, they often can be tentative. It is important that we keep them so.

References

1. Davis, Frederick B. "Research in Comprehension in Reading," *Reading Research Quarterly*, 3 (Summer 1968), 499-545.
2. Kelley, Truman L. *Interpretation of Educational Measurements*. New York: World Book, 1927.

Mary M. Brittain
University of Wisconsin

GUIDELINES FOR EVALUATING CLASSROOM ORGANIZATION FOR TEACHING READING

The variety of classroom organizational patterns for reading instruction is enormous. Even a limited sampling of the organizational "smorgasbord" renders one intellectually replete: total class grouping arrangements (whether on a temporary basis as for choral reading or open text sessions, or on a more permanent footing as in tracking or special reading classes); cross-class groupings, such as in the ungraded primary or Joplin approaches; intraclass grouping, including bi-, tri-, or multibasal grouping, grouping by invitation, grouping to meet special interests or skill needs of pupils, student-led small team grouping, tutorial grouping. The spectrum extends to complete individualization of instruction.

These multitudinous organizational patterns are all attempts to increase the teacher's efficiency in meeting the reading needs of individual children, and most, though not all of them, seek to do this through the reduction of pupil heterogeneity. (Actually, a number of the plans share many other common attributes, a factor which complicates the evaluative process.) While it is not feasible here to undertake any extensive comparison of organizational patterns, it is perhaps possible to suggest a combination of evaluative approaches that will render such comparisons meaningful.

EVALUATION THROUGH STANDARDIZED TESTS

In the main, the evaluation, not only of classroom organization, but of most of the elements of the reading program, has been done in terms of their impact on student growth (*5*) and growth has been measured most frequently by means of standardized

tests. The inadequacies of standardized tests as measures of pupil growth are well known, but it may be helpful to note some of the distinctions between the processes of testing and evaluation, and to question the assumption that student growth in reading skill should form the sole basis for evaluation.

Ammons (*1*) has defined evaluation as the "description of student progress toward educational objectives," and has noted that, in contrast to testing, evaluation is directed more to individuals than to groups, and seeks to describe the progress of the individual student toward certain school-defined objectives. Standardized tests, on the other hand, are not typically criterion-referenced; rather they aim to compare the progress of a group (less successfully that of an individual) with that of other (normative) groups and, while these tests may indicate the level of a group's performance, they seldom provide insights as to why a group performs as it does.

Standardized measures have some further shortcomings as evaluative instruments. If we assume that evaluative procedures should be ongoing and should provide sufficient examples of a student's work to sample the various skills of reading adequately, then any one-shot temporally-discrete testing method will be found wanting. The importance of repeated sampling in evaluation can scarcely be overemphasized, significant differences in test performance having been demonstrated with changes in examiner, test content, physical setting, and time of day or year.

If standardized measures are to yield any relevant information, care must be taken to select tests that are appropriate to the content of the reading program—tests that actually measure the behaviors deemed important in accomplishing the school's objectives. Testing instruments may, in fact, bear little relation to the objectives of a particular educational program. For example, the use of a typical reading achievement test to evaluate an organization whose primary objective is the development of more positive attitudes toward reading would be an exercise in futility.

GOAL-REFERENCED EVALUATION

If, as Barrett (*3*) has suggested, instructional goals should form the basis of evaluation, then the philosophy of the school regarding reading, and the manner in which the school defines the reading process assume importance. While it is certainly possible to

define reading differentially, it will be assumed for the purposes of this paper that reading is not solely a perceptual or cognitive process (though it certainly subsumes these elements), but includes affective aspects such as appreciation and enjoyment. This being the case, evaluation of any organizational strategy for reading instruction may well begin, as Russell and Fea (*9*) have suggested, with an analysis of the characteristics of successful readers in all the above-mentioned parameters of reading—their habits, skills, attitudes, and interests—and the behaviors that are implicit in the development of these characteristics. Such an analysis should yield hypotheses regarding the ideal classroom arrangement for evoking the desired behaviors. For example, if one assumes that the successful reader is characterized by concentration on the meaning of a selection, one may hypothesize that solitude while reading, free from group distractions, would be conducive to the development of concentration and therefore opt for an individualized approach. However, one may also conceptualize the good reader as one who can successfully interpret a selection for the pleasure or profit of others. One might then suppose that small-group organization for oral reading would be desirable. Evaluation would proceed in terms of how successfully the organizational pattern promoted the skills or attitudes enumerated in advance—in effect, a criterion-referenced approach. The evaluation should also include a statement of which features of the classroom organization plan appeared to contribute to which outcomes.

Research on the effectiveness of grouping for reading instruction is neither copious nor consistent, perhaps partly because of the application of inappropriate measures, but also because of the failure of research studies to include sufficiently detailed descriptions of the instructional practices employed, that is, the implementation of the organizational procedures. The conflicting results from the USOE first grade studies of reading instruction (*4*) offer ample testimony to the difficulty of being sure of what one is actually evaluating.

Classroom organizational patterns for teaching reading, whatever their particular form, must answer to the demands inherent in the nature of reading, those inherent in the learner, and those inherent in the resources of the school. In an effort to provide some useful guidelines for evaluation, certain organizational characteristics may be hypothesized as favoring the fulfillment of

these multiple demands. The following organizational standards have been so derived.

Goal direction

Given the non-unitary character of the reading process, which presupposes a multiplicity of instructional goals, evaluation of classroom organizational procedures (whether the classroom is self-contained or of the multi-unit sort) must include an estimate of their efficacy in promoting the various aspects of reading. Further, grouping strategies should be examined for their facilitation of growth toward expressly stated goals. Some examples follow:

Word Perception—Does the organization promote flexibility in methods of word attack?

Comprehension—Does the organization lead to inferential and critical responses to what has been read?

Appreciation—Does the organization facilitate responses to artistic, humorous, or stylistic elements of selections?

Rate of Reading—Does the organization foster flexibility of reading rate?

Oral Reading—Does the organization develop skill and enjoyment in oral reading?

Study Skills—Does the organization advance growth in reading specific to the various content areas?

Flexibility

For many years it has been suggested that good organizational strategy should provide for flexibility of group size and membership, that students should be afforded the opportunity to work not only with the whole class, but with small groups and individuals. In view of the complexity of the reading process and the probability that different organizational arrangements will be conducive to differential skill development, the criterion of flexibility would appear sound. A further advantage of group flexibility is that learners have the opportunity to work with others who may or may not be similar in general attainment, but who share common skill development needs or common interests. Moreover, flexibility of grouping would reduce the likelihood of a stigma being attached to perpetual membership in a "low" group, thus contributing to the healthy development of the self-concept.

Interests

Current statistics relating to the reading habits of adults (2) demonstrate all too clearly that instructional programs in the nation's schools could be greatly improved insofar as the promotion of reading as a leisure-time pursuit is concerned. Concern for the affective aspect of reading suggests that the organizational plan adopted should not only allow for pupils' self-selection of materials and self-pacing in these, but should permit the extension of interests through teacher-pupil and pupil-pupil exchanges.

Independence

An important goal of reading instruction is to produce mature readers who are able to gain both pleasure and profit from printed material without the constant assistance or direction of a teacher. Data from transfer of training studies (7) suggest that self-direction in reading activities should receive an early introduction. Supplementation of teacher-directed groups with those directed by individual students should also increase pupil awareness of program goals through greater involvement in planning and implementation.

Homogeneity

While heterogeneity is a healthy fact of group life, any grouping, whether by ability, achievement, or interest, should be sufficiently homogeneous to afford a reasonable opportunity of success or self-fulfillment to the members. This is not to suggest that rigid, narrow criteria for group membership should be established, but requiring an individual's membership in a group that is grossly divergent from him in needs, preferences, and attainments can hardly be justified on either cognitive or affective grounds.

Instructional Personnel

The organizational plan should be realistic in terms of the degree of teacher expertise required and should be mindful of individual differences between teachers as well as pupils. For example, it has long been noted that individualized reading programs, while extremely effective in improving children's attitudes toward reading, require teachers of notable independence and competence. A teacher with more modest attainments, or one who is more secure within a structured framework, would probably

function more effectively under an alternative arrangement. Teachers should be able to select, from a number of organizational patterns, those that enhance, rather than inhibit their effectiveness.

Administration

Certain logistical problems must also be addressed. Do the proposed grouping procedures promote ease of scheduling? Do they require resources whether of materials, space, or personnel that are within the school's capacity?

A checklist based upon the foregoing characteristics of grouping practices is appended to this paper. The checklist suggests important areas of concern relating to the learner, the reading process, and administrative concerns and, within these categories, includes sample questions that may guide classroom organizational patterns for teaching reading.

ALTERNATIVE EVALUATIVE PROCEDURES

If it is assumed that, given the complexity of the reading process, a variety of evaluative techniques are requisite, what supplements to standardized measures are available? It must be admitted at the outset that there exists no single well-established theory of methodology for measuring classroom behavior, but some reasonable possibilities—not without their own limitations as to reliability and relevance—include:

> Repeated systematic observation within the classroom through such media as film, kinescope recordings, observers utilizing rating scales. (If, for example, one wishes to assess interest in reading as a leisure-time pursuit, observation of free-choice situations in which students may select from a variety of activities should provide valuable insights regarding level of interests in recreational reading.)
>
> Paper and pencil measures, such as anecdotal records kept by teachers and pupils relating to types and amounts of reading done; records of comprehension difficulties; vocabulary files; interest and attitude inventories; social adjustment measures; teacher-made checklists of reading skills.
>
> Informal estimates of reading habits, skills, attitudes, and interests such as may be derived from performance on informal reading inventories, performance in content area reading, tapes of students' oral reading, records of student library usage, and of out-of-school reading habits. Students and parents, as well as teachers, may contribute to these informal evaluations.

CONCLUSION

The aim of this paper has been to present a theoretical framework for the evaluation of organizational patterns for reading instruction and to suggest some supplementary approaches to the traditional use of standardized measures. Of necessity, the treatment has involved sampling from many different aspects of reading instruction, since any organizational strategy exists chiefly to advance the many and complex goals of reading instruction. It is on the quality of service to all these masters that evaluation of organizational patterns must proceed.

A Checklist for Evaluating Classroom Organization for Teaching Reading

STUDENT CHARACTERISTICS

Physiological

Do pupils have sufficient opportunities for movement?
Are special sensory needs of pupils met?

Social

Are students' group roles clearly defined?
Is student-direction of learning situations encouraged?

Affective

Is a reasonable opportunity of success ensured?
Is stigmatization avoided?
Can the varied interests of students be met?

Cognitive

Are the experience backgrounds of students utilized?
Can differing rates of learning be provided for?
Are pupils able to exchange opinions regarding selections?

Educational

Are planned experiences to meet specific skill needs possible?
Are sufficient opportunities provided for diagnosis?

INSTRUCTIONAL GOALS

Word Recognition

Are a variety of word recognition methods practiced?
Is accuracy of word perception facilitated?

Comprehension

Are inferential as well as literal skills facilitated?

Are exchanges of critical judgments encouraged?

Appreciation

Are students' reading interests broadened?

Are students' reading attitudes and habits improved?

Rate of Reading

Is flexibility of rate encouraged?

Is pressure to maintain a group standard of rate avoided?

Oral Reading

Is a reasonable balance maintained between oral and silent reading?

Are meaningful alternatives to round-robin reading provided?

Content Area Reading

Are the various study skills practiced?

Is there opportunity for students to apply what they have read?

IMPLEMENTATION

Teacher Personnel

Is teacher expertise maximally utilized?

Are teacher preferences and interests considered?

Data Collection

Can adequate samples of students' reading behavior be obtained?

Can information regarding growth in specific skills be obtained?

Scheduling

Is scheduling simplified?

Can flexibility of instructional time be maintained?

Materials

Can a variety of materials be employed?

Are the requisite materials within the school's financial resources?

References

1. Ammons, Margaret. "Evaluation: What Is It? Who Does It? When Should It Be Done?" in Thomas C. Barrett (Ed.), ***The Evaluation of Children's Reading Achievement.*** Newark, Delaware: International Reading Association, 1967, 1-12.
2. Asheim, Lester. "What Do Adults Read?" ***Fifty-Fifth Yearbook of the National Society for the Study of Education,*** Part II. Chicago: University of Chicago Press, 1956, 5-28.

3. Barrett, Thomas C. "Goals of the Reading Program: The Basis for Evaluation," in Thomas C. Barrett (Ed.), *The Evaluation of Children's Reading Achievement*. Newark, Delaware: International Reading Association, 1967, 13-26.
4. Bond, Guy L., and Robert Dykstra. "The Cooperative Research Program in First Grade Reading Instruction," *Reading Research Quarterly*, 2 (Summer 1967), 5-142.
5. Farr, Roger. *Reading: What Can be Measured?* Newark, Delaware: International Reading Association, 1969.
6. Heathers, Glen. "Grouping," in Robert L. Ebel (Ed.), *Encyclopedia of Educational Research* (4th ed.). New York: Macmillan, 1968, 559-570.
7. Klausmeier, Herbert J. "Transfer of Learning," in Robert L. Ebel (Ed.), *Encyclopedia of Educational Research* (4th ed.). New York: Macmillan, 1968, 1483-1493.
8. Medley, Donald M., and Harold E. Mitzel. "Measuring Classroom Behavior by Systematic Observation," in N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963, 247-328.
9. Russell, David H., and Henry R. Fea. "Research on Teaching Reading," in N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963, 865-928.

Morton Botel
University of Pennsylvania
John Dawkins
Research for Better Schools
Philadelphia, Pennsylvania
Alvin Granowsky
Diagnostic Reading Center
Greensboro, North Carolina

A SYNTACTIC COMPLEXITY FORMULA

A reliable and valid measure of complexity of syntactic structures is of theoretical interest and should be helpful in preparing and selecting reading materials. A count of the number of words per sentence, the measure most widely used at this time, has been judged inadequate by newer theories of grammar as well as by research findings. Other generally used measures of syntactic complexity focus only on a few syntactic structures which correlate somewhat with reading complexity, but in no way indicate the relative complexity of the major portion of syntactic structures found in reading materials.

In an attempt to overcome these inadequacies, a heuristic was developed—a syntactic complexity formula (*1, 2*). This formula is based on 1) a theory of transformational grammar that suggests that complex sentences can be thought of as derived from processes of changing and combining underlying structures (simple sentences, for our purposes); 2) experimental data on children's processing of syntactic structures; and 3) language development and performance studies of the oral and written language used by children.

A measuring device that takes into account multiple factors of syntax will reveal a great deal of information about what is and what is not hard to process for the young reader. However, the device will also have limitations that must be mentioned. First, there are a number of factors in syntax, and many factors in semantics, that do not readily lend themselves to measurement.

Second, there are small degrees of differences in syntactic difficulty that cannot be rated on a scale without making it far too cumbersome to use. For this reason, we have rated many items as equivalents, when some differences in their complexity clearly exist.

Finally, two cautions need to be noted in using the syntactic complexity formula: 1) It should be used in conjunction with a measure of vocabulary; and 2) the value of the instrument lies not in giving a precise measurement but in ranking syntactic structures.

To apply the syntactic complexity formula to any passage, each sentence in the passage is assigned a complexity rating. These ratings are then averaged to obtain the complexity rating for the entire passage. The complexity rating for a sentence is determined by comparing the structure of the sentence to the structures described and illustrated on the following pages. The basic structure of the main clause of the sentence is assigned a count of 0, 1, or 2 and counts are added for additional features or structures that add complexity. For example, the sentence ***His vacation over, the tired doctor drove home*** has a complexity count of 4: The basic structure SV(Adv) (***The doctor drove home***) gets a count of 0 (see IA under "0-Count Structures"). Since the subject (***doctor***) is modified by an adjective (***tired***) a count of 1 is added (see IIIA under "1-Count Structures"). The absolute (***His vacation over***) at the beginning of the sentence adds an additional count of 3 (see II under "3-Count Structures"). The whole sentence thus receives a count of 0+1+3 = 4.

0-Count Structures

I. The Most Frequently Used Simple Sentences		Count
A. ***Subject-Verb (Adverbial)***	SV(Adv)	
He went.		0
Bob had gone.		0
The boy had gone.		0
That boy had gone (home).		0
Those girls have been playing (at their house).		0

B.	*Subject-Verb-Object*	SVO	Count
	She hit it.		0
	The fish weighed a pound.		0
	Those girls have been hitting my ball.		0
C.	*Subject-Verb-be-Complement**	SbeC	
	pattern 1: adjective	SbeC-adj.	
	He is big.		0
	He is very big.		0
	The girl seemed big.		0
	These children will have grown big.		0
	pattern 2: noun	SbeC-noun	
	She became president.		0
	That girl was their president.		0
	Those students have been our presidents.		0
	pattern 3: adverbial	SbeC-adv.	
	He is there.		0
	It is there.		0
	She will be there.		0
	Those girls have been in their homes.		0
D.	*Subject-Verb-Infinitive*	SVInf.	
	Bob wanted to go.		0
	These girls will want to eat.		0
	Our children have been waiting to eat.		0

II. Simple Transformations

A. ***Interrogative***

1. ***Simple question:***

Will he run?	0
Did he do it?	0

2. ***Tag-end question***

Declarative sentences can become questions by adding sentence tags:

The game was good, wasn't it?	0
He did it, didn't he?	0

B. ***Exclamatory***

What a game!	0
What a game it was!	0
How wonderful!	0

C. ***Imperative***

(You) Get the milk. (!)	0
(You) Go to the store.(!)	0

*Linking verbs such as ***seem, became, turn,*** are included in the category of "be" verbs.

III. Coordinate Clause Joined By *And*

Research indicates that the coordinate clause joined by *and* represents one of the most common, easily processed structures in the language.

	Count
John went to the store.	0
Mary went to the store.	0
John went to the store ***and*** Mary went to the store.	0

IV. Nonsentence Expressions

1.	noun of direct address:	Is that you, MARY?	0
2.	greetings:	Hi, Hello	0
3.	calls and attention getters:	Hey	0
4.	interjections:	What, Wow, Oh	0
5.	responses:	Okay, Good-by, So long	0
6.	empty phrases:	Really now, You know	0
7.	sentence openers:	Please, Then, But then	0

Note:

I. There is no extra count for these expansions of simple sentences.

A. Verb expansions: 1. forms of ***be, have*** and ***do***
2. ***will*** and ***can***

B. Intensifier expansions: ***very, too, so, much, more*** even when two are used together as in ***much more, so very,*** etc.

C. Determiner expansions:
1. articles: ***a, an, the***
2. demonstrative pronouns: ***this, these, that, those***
3. possessive pronouns: ***my, our, your, his, its, their***

1-Count Structures

I. Two Less Frequently Used Sentence Patterns

	Count
A. ***Subject-Verb-Indirect Object-Direct Object*** SVIO	
He threw HER the ball.	1
B. ***Subject-Verb-Object-Object Complement*** SVOC	
They made him HAPPY.	1

II. Any Prepositional Phrase Added to Any 0-Count Pattern

		Count
A.	*Subject-Verb (Adverbial)* *	
	The boy had gone home IN THE MORNING.	1
B.	*Subject-Verb-Object*	
	The girl threw the ball TO THE CATCHER.	1
C.	*Subject-be-Complement*	
	The man BEHIND THE DESK was big.	1
D.	*Subject-Verb-Infinitive*	
	Bob wanted to go BEFORE BILL.	1
III.	Noun Modifiers	
A.	*Adjectives*	
	The BIG man ate here.	1
B.	*Nouns*	
	Their team ate the APPLE pie.	1
C.	*Predeterminers* (one of, two of, many of, both of)	
	ALL OF the players won the game.	1
D.	*Possessive Nouns*	
	The hat fit his SON'S head.	1
E.	*Participle* (*ed* and *ing* forms in the natural adjective position)	
	The CRYING boy ran home.	1
	The SCALDED cat ran home.	1

GENERAL RULE:

A 0-Count sentence has three or fewer lexical words.
A 1-Count sentence generally has four lexical words.

Lexical words are nouns, verbs, adjectives and adverbs. Prepositions are not counted as lexical words. In general, each is given a 1-count when added to a basic sentence pattern.

IV.	Other Modifiers	
A.	*Adverbial Additions to the 0-Count Sentence*	
	He ran to the store LATER.	1
	He QUICKLY went to the store.	1
B.	*Modals*	
	(could, dare to, has to, may, might, must, need to, ought to, shall, should, would)	
	He MIGHT have won the game.	1

*The first adverbial in a subject-verb (adverbial) pattern is not given a count.

	Count
C. *Negatives* (no, not, neither, never, n't)	
He did NOT see it.	1
He didN'T do it, did he?	1

V. Set Expressions

These are phrases that are usually strung together. They are given a 1-count, even if their lexical number is higher than one.

Many years ago, Once upon a time, (Every) once in a while, (Every) now and then, a_____year old (modifier), _____years old (complement), more or less, etc.

VI. Infinitives

When the infinitive does not immediately follow the verb, it is considered an expansion of the basic sentence pattern and given a count.

They wanted the baby TO SLEEP.	1
They tried hard TO REST.	1

VII. Gerund

When the gerund is a subject, it is given a count. (In all other uses, the gerund is counted as any other noun.)

RUNNING is fun.	1

VIII. Coordinate Clause (joined by coordinate conjunctions other than and: *for, but, so, yet, or*)

John worked hard.	0
He played hard.	0
John worked hard BUT he played hard.	1
The boy did that job OR you did it.	1
The BIG boy did that job OR you did it.	1 + 1 = 2

IX. Deletion in Coordinate Clauses

This process is already accounted for by the "General Rule" on lexical additions. Note that *and* is included here.

John was thin.	0
John was healthy.	0
John was thin *but* HEALTHY.	1
Joe jumped into the water. Pete jumped into the water.	0 + 0
Joe *and* PETE jumped into the water.	1
Joe *and* HIS FRIEND PETE jumped into the water.	1 + 1 = 2

X. The Paired Conjunction *both . . . and*

BOTH Bob did it AND BILL did it.	1
BOTH Bob AND BILL did it.	1 + 1 = 2

2-Count Structures

	Count
I. Passive Transformations	
The ball was hit by Bob.	2
The ball was hit. (by Bob, understood)	2
II. Paired Conjunctions (either . . . or, neither . . . nor, not . . . but, etc.)	
NEITHER Pete did it NOR Bill did it.	2
When deletion is involved, simply count the lexical items.	
NEITHER Pete NOR BOB did it.	2 + 1 = 3
NEITHER Pete NOR *my* FRIEND BOB did it.	2 + 1 + 1 = 4
III. Comparatives as_____as; same _____as; er_____than; more_____than	
Bob was AS tall AS Bill (is).	2
She is MORE attractive THAN you (are).	2
IV. Dependent Clause	
A. *Adjective clauses*	
The book (THAT) I READ was great.	2
The postman, WHO DELIVERS THE MAIL, is nice.	2
B. *Adverbial clauses*	
He left WHEN HE FINISHED.	2
He came early SO THAT HE COULD BUY THE GIFT.	2
C. *Nominal clauses*	
He asked me WHAT I DID.	2
V. Participle	
When attached as a modifier in a typical adjective-noun order: 1-count. But when the participle appears after the noun or is separated from it by commas, give it a 2-count.	
BOILING, the water overflowed the pan.	2
The water, BOILING, overflowed the pan.	2
YOWLING, the *scalded* cat ran home.	2 + 1 = 3
VI. Infinitive as Subject	
TO RUN is healthy.	2
VII. Appositive	
To be considered a 2-count appositive, the structure must be a noun phrase set off by commas.	
His good friend, a pretty girl, arrived.	4
(adjectives: good, pretty = 2-count appositive: a girl = 2-count)	

VIII. Conjunctive Adverbs

Examples: thus, moreover, however, therefore, consequently, nevertheless (and also *still* and *yet* when used as conjunctive adverbs).

	Count
I went, NEVERTHELESS.	2
YET, everyone applauded.	2

3-Count Structures

	Count
I. Clauses Used as Subjects	
THE FACT THAT HE EATS is important.	3
THAT HE EATS is important.	3
II. Absolutes	
THE JOB FINISHED, Bob went home.	3
Mr. Smith lit his pipe, THE PERFORMANCE OVER.	3

Special Handling

I. Noun Clause of Dialogue

Procedure for counting: Separate the speaker from what is said and count the parts as two sentences.

(1) John said, "I will go." = 0-count
  (a) John said. = 0-count
  (b) I will go. = 0-count

If either part carries a count, consider it as you would any sentence:

(2) The big bird chirped, "Go away!" = 1-count
  (a) The big bird chirped. = 1-count for adjective
  (b) Go away! = 0-count

Structures similar in format to the Noun Clause of Dialogue, such as ***say, wonder, believe, feel,*** will be handled in the same manner.

(3) I wondered who would do it. = 1-count

(a) I wondered. = 0-count

(b) who would do it. = 1-count for modal

(4) Those terrible boys who live on our street said that we should go. = 4-count

(a) Those terrible boys who live on our street said = 1-count for the adjective, 2-count for the adjective clause, total = 3-count.

(b) (that) we should go. = 1-count for the modal

II. Inverted Order of Adverbials of Manner and Place

Whenever these adverbial structures begin the sentence, add a 1-count to the scoring you would typically give:

	Count
He ran to the store QUICKLY.	1
QUICKLY, he ran to the store.	2

III. Names and Titles

Names and titles, whatever their length, should be regarded as a simple noun in scoring.

MR. WILLIAM JONES is here.	0
THE AMERICAN RED CROSS helps people.	0

IV. Hyphenated Words: Count as Separate Words, If the Parts Can Stand Alone

The never-ending day is never ending.	3

PROCEDURE FOR DETERMINING AVERAGE SYNTACTIC COMPLEXITY

The syntactic complexity of any passage or sampling of sentences is the arithmetic average of the complexity counts of the sentences evaluated. For example, if ten sentences had the following counts, their average syntactic complexity would be 2.5.

1. 2	6. 2	
2. 2	7. 1	
3. 3	8. 4	total 25
4. 1	9. 3	average 2.5
5. 2	10. 5	

PROGRAMING SYNTACTIC COMPLEXITY

Syntactic complexity of reading materials may be graded from a starting point of 0-count complexity to any *average* syntactic complexity count designated a terminal reading level.

For example, syntactic complexity of materials prepared for a primary reading program may begin at the 0-count level and progress to an average complexity count of 3.0 to 4.0.

Application of the formula is shown in the paragraph analyzed below:

Daedalus, the First Man to Fly

(1) Daedalus jumped from the mountain top. (2) For a terrible moment, he fell straight down, his arms wobbling weakly. (3) But then he spread his wings and began to fly. (4) Like a bird, he flew straight up into the blue morning sky.

	0-Count	1-Count	2-Count	3-Count	Total
1.	SVAdv.	adjective			1
2.	SVAdv.	prep. phrase adjective adverb		absolute	6
3.	SVO	coord. clause deletion: two lexical items			2
4.	SVAdv.	prep. phrase adjective adjective	prep. phrase: inverted order		5
				Total	14

14 divided by 4 = 3.5 = average syntactic complexity.

References

1. Botel, Morton, and Alvin Granowsky. "A Formula for Measuring Syntactic Complexity: A Directional Effort," *Elementary English,* 49 (April 1972), 513-516.
2. Granowsky, Alvin. "A Formula for the Analysis of Syntactic Complexity of Primary Grade Reading Materials," unpublished doctoral dissertation, University of Pennsylvania, 1971.

Theodore A. Mork
Western Washington State College

THE ABILITY OF CHILDREN TO SELECT READING MATERIALS AT THEIR OWN INSTRUCTIONAL READING LEVEL*

During the past several years, much emphasis has been placed on programs involving self-selection of reading materials. Special library-centered reading programs expect children to select materials that are appropriate for them in terms of interest, maturity level, and reading difficulty. Authors of basal readers emphasize the importance of children's reading library books in conjunction with their basic texts, and, for the most part, the selection is left up to the children. In individualized reading programs based on self-selection of reading materials, an ability to select materials of appropriate difficulty is essential (*1,4,6,8,18*).

However, while several educators have emphasized that children must be allowed to select their own reading materials, the ability of children to choose materials appropriate to their reading abilities appears largely to have been assumed.

During the past four or five years, many university students in education and many practicing teachers have seriously questioned this assumption. Others have wondered how much guidance or help might be necessary to increase children's ability to select materials of appropriate levels of difficulty. Jacobs (*7*) suggested: "The teacher will probably have to give some guidance, helping the child to be realistic about his choices in terms of his capabilities, his aspirations, his past experiences." Vite (*19*), speaking from a primary teacher's experience, also suggested that the child himself selects, but with guidance and support from the teacher as needed.

*This study was supported by the Educational Research Institute of British Columbia.

The lack of information available on the difficulty level of books that children actually do select prompted this research study. Its purpose was to provide empirical information about some practical aspects of self-selection that may be helpful to teachers and to teachers of teachers. Specifically, the study sought to answer the following questions: 1) In relation to a child's instructional reading level, what level of materials, particularly library books, does a child select when he is allowed freedom to make his own choice? Do children, in fact, select materials of appropriate difficulty without guidance from the teacher? 2) Does a short, five-minute period of guidance from the teacher that emphasizes self-acceptance affect the child's selection?

In addition, the relationships between the observed discrepancy scores (the difference between instructional reading level and the level of materials chosen) and sex, age, and reading ability were explored.

SUBJECTS

Twenty-nine children in grade three and thirty-one in grade five were randomly selected from a group of 200 children in Victoria, British Columbia, elementary schools. These 200 children had already been selected through random procedures for a research study being conducted by Tinney (*16*). The children were selected from eight different schools, representing a cross-section of elementary school children. In the present study no more than eight children were selected from each school, and no more than four children were selected from each grade level (third or fifth) in each school. In all but two cases, the children came from different classrooms. According to the building principals, none of the children in the study had been involved in an individualized reading program for their reading instruction.

PROCEDURES

Each child in the study was asked to select, from each of three different sets of reading materials, a piece of material that he thought he could read fairly well by himself, with perhaps some occasional help from the teacher. This, essentially, was a definition of instructional reading level as interpreted for the child. The three sets of reading materials were: 1) single pages copied from basal readers, 2) a series of basal readers, and 3) library books.

The children at each grade level were randomly divided into two groups. Children in the ***guidance group*** were engaged in discussion relative to the differences found in children, such as running speed, height, weight, and shoe size. A comparison was drawn between shoe size and "book size" (*10*).

The intent of this brief session was not to "tell" the child anything, but rather to lead him through a sequence of questions and answers that would help him to conclude that it was normal for children to differ in their reading abilities and that selecting materials for "reading fit" was rather similar to finding clothing of the right size. Except for this brief session, the tasks of the children in the two groups were identical.

Each child met with an examiner in an individual session at the child's school. If the child was in the guidance group, the examiner first discussed with him the nature of individual differences and "reading fit" as described above. The first selection task for children in both groups involved separate pages that had been randomly selected and photocopied from the first half of basal readers at successive levels of difficulty, preprimer through eighth grade. These readers were from a series unfamiliar to the children (*14*). The child was asked to select, from five of the pages, the page that he could read fairly well, but with which he might need just a little help from the teacher. The five pages included the page from the reader for the child's grade level and extended two levels in each direction. If the page chosen was the lowest or the highest difficulty level of the five, the child was asked to look at an even lower or an even higher level before making his final choice. If necessary, he was shown additional levels as well. The difficulty level of the material finally selected was recorded.

Next, the child was asked to perform the same task with five books from the same series of basal readers. Interest in any single story may have affected his selection; thus, once the child had made a tentative selection, the examiner suggested that he look at several different stories in the selected reader before making a final choice. Again, if the lowest or highest level reader was selected, the examiner suggested that the child look at an even lower level reader or an even higher level reader before making his final choice.

The final step was for the child to go to the school library, which was a larger source than most classroom libraries, and to select a book to read using the same criteria described above. A

readability check was run on the book to determine the difficulty level.

The instructional reading levels for these children had been established for a separate study (*16*). A reading inventory had been administered by advanced students in Reading Education at the University of Victoria. The reading inventory used required the child to read one passage orally and one silently at each successive reader level. Oral reading errors were scored, as were oral and silent reading comprehension. The instructional reading level was the highest level at which the child could read with at least 95 percent word recognition and at least 70 percent comprehension. The criteria for determining the instructional levels were based on those suggested by Betts (*2*). The examiners in the present study were not informed of the instructional reading levels of the children until after the children had made their selections.

The difficulty levels of the separate pages and of the basal readers were substantiated using the Fry Readability Graph (*5*). Readability was checked to be sure the chosen pages were representative of the difficulty of the books from which they were selected. The readability levels of the library books chosen by the children were also determined using the Fry graph.

The differences between the established instructional reading levels (grade scores) and the difficulty levels (grade scores) of materials selected by the children were determined. These differences (*discrepancy scores*) for the guidance group and the no-guidance group were then compared.

RESULTS AND DISCUSSION

It appeared that for many of the children, trials one and two—selecting a separate page and selecting a basal reader—functioned as a training situation. That is, several of the children seemed to be learning the task during these two situations. For this reason, the comparisons between groups were based on the discrepancy scores for the library book selections. An additional reason for using the library book selections is that library books represent more accurately the type of material with which self-selection normally occurs in the elementary classroom.

The effect of the five-minute period defined as guidance was evaluated by a *t*-test of difference between the mean discrepancy

scores of the guidance and the no-guidance groups. The means and standard deviations of the discrepancy scores are reported in Table 1.

TABLE 1

Means and Standard Deviations of Discrepancy Scores for Selections

Group		Based on Signed Values[a]	Based on Absolute Values
Guidance	Mean	-0.63	1.33
	S.D.	1.55	1.02
No Guidance	Mean	-0.32	1.12
	S.D.	1.57	1.14

[a]Positive value means instructional level higher than readability level of selection.

Note that means and standard deviations are reported in two ways, first using their signed (positive or negative) values, then their absolute values. Because the positive and negative discrepancy scores tended to offset each other, these means of the signed values are somewhat deceiving. The means of the absolute values give a more accurate picture of the actual distances between instructional reading levels and the levels of material selected. The difference between the guidance and the no-guidance groups was not significant for either the signed or absolute discrepancy scores, and it must be concluded that the guidance had no effect. Whether regularly repeated guidance of this or another sort might be effective remains to be studied.

After it had been established that no significant difference existed between the guidance and the no-guidance groups, the data from all subjects were combined to determine the level of materials children actually do select relative to their instructional level. A *t* test for matched groups was used to evaluate the difference between the children's instructional reading levels and the difficulty levels of their library book selections. The mean instructional reading level (4.77) for the 60 children was significantly lower than the mean difficulty level (5.24) of their selections ($t = 2.33$, $p < .05$). Thus, the children did not, as a group, select materials with difficulty level equal to their established instructional reading levels but tended to select material at a higher level.

However, the actual difference between the mean instructional reading level and the mean difficulty level of the library books selected was less than one-half year.

How large a difference can be tolerated between a child's instructional reading level and the difficulty level of the material he selects, and still have the material appropriate for him? McCracken (*9, 10*) contends that a child's instructional reading level usually is a range of two or more book levels. We know that children's interests often allow them to read materials otherwise thought to be too difficult for them. The data were examined to see how many children had discrepancy scores on their library book selections of one year or less, that is, chose a library book within one year of their instructional reading level. For this analysis, the children were placed into groups of high, middle, or low reading ability on the basis of their instructional reading levels in relation to their grade levels. The study took place in April. Therefore, in grade three, children whose instructional reading levels were greater than 3.5 were considered high in reading ability, those between 3.0 and 3.5 were classified as middle reading ability, and those less than 3.0 were placed in the low reading ability group. In grade five, the corresponding high, middle, and low reading ability groups were those whose instructional reading levels were greater than 5.5, between 5.0 and 5.5, and less than 5.0.

TABLE 2

Number and Percentage of Subjects Selecting Library Books with Difficulty Levels Within One Year of their Instructional Reading Levels

Reading Ability	Grade 3		Grade 5		All Subjects	
	No.	%	No.	%	No.	%
High	10/16[a]	63	7/9	78	17/25	68
Middle	4/5	80	9/15	60	13/20	65
Low	2/8	25	5/7	71	7/15	47
Total	16/29	55	21/31	68	37/60	62

[a]Read as 10 out of 16

It can be seen from Table 2 that 62 percent of all the subjects selected library books with difficulty levels within one year of

their established instructional reading levels. The percentage of children attaining the criterion of plus or minus one year increases as reading ability increases and is somewhat higher for the older children. It is important to note that the children in grade three whose instructional reading levels are below grade level did least well in selecting appropriate library books. Only two out of eight selected material within one year of their established reading levels. Of the 37 children who selected library books within one year of their instructional levels, 19 were female, 18 were male, 21 were in grade five, and 16 were in grade three.

Eighteen children obtained a discrepancy score of zero for their library book selections. This was nearly one-third of the subjects. Seven were male, eleven were female. Ten were fifth graders, eight were in third grade. Of the eight in grade three, six were in the high reading ability group with one each in the middle and low groups. Of the ten subjects in grade five, five were in the high group, three were in the middle group, and two were in the low group. These figures suggest that the higher the reading ability, the more likely a child will be to make appropriate book selections in terms of difficulty levels.

CONCLUSIONS

The following conclusions appear warranted by the results of this study.

1. A five-minute period of individual guidance as defined in this study will not influence a child to select reading materials more appropriate to his instructional reading level. It is possible, however, that used over a period of weeks or months, this approach might be successful.

2. Many children are able to select reading materials that are exactly the same as their instructional reading levels as determined by informal reading inventories. Nearly one-third of the subjects obtained a discrepancy score of zero for their library book selections.

3. If it is accepted that materials appropriate for a given child range from one year below to one year above his instructional reading level, then the majority of children in grades three and five can choose appropriate books but a substantial minority will need guidance in making selections. More than 60 percent of the children selected materials appropriate to their reading levels, but

nearly 40 percent did not.

4. On the whole, older children and better readers appear somewhat more able to select reading materials of appropriate difficulty.

LIMITATIONS OF THE STUDY AND RECOMMENDATIONS FOR FURTHER RESEARCH

The results of this study should be interpreted in light of certain limitations. Some of these limitations and some of the positive findings suggest profitable further research.

Use of an informal reading inventory could have allowed for error in determining the base level from which to make comparisons. Use of a standardized reading inventory such as the Spache Diagnostic Reading Scales (*15*) or the Standard Reading Inventory (*9*) might have provided increased accuracy.

The role interest played in the selection of library books was not controlled. It is generally accepted that the child's interests and experiential background may cause him to choose a book that is somewhat above his general instructional reading level but that his interest and knowledge of specialized vocabulary may allow him to read profitably. Evaluation of how well the child could read the specific material he selected would carry the investigation a step farther and provide some information on the effect of interest and specific subject matter on the child's instructional level. For this evaluation, the standard criteria for assessing difficulty of materials through oral reading (*2*) could be employed, or the cloze procedure (*3*) could be used. In a future study, also, having each child make more than one library book selection would provide additional confidence in results.

According to the building principals, the subjects of this study had never been exposed to an individualized reading program. The majority of their reading instruction had been from basal readers. Even though most basal reading programs do make recommendations about the use of library books, including some suggestions on self-selection, little emphasis had been placed on helping these children to select appropriate reading materials. It would be particularly interesting to repeat this study with a group of children who, for several months, had been involved in a reading program based on self-selection. A brief longitudinal study could clarify the effects of continued guidance and practice on the ability of children to select materials of appropriate difficulty.

References

1. Barbe, Walter B. ***Educator's Guide to Personalized Reading Instruction.*** Englewood Cliffs, New Jersey: Prentice-Hall, 1967.
2. Betts, Emmett. ***Foundations of Reading Instruction.*** New York: American Book, 1946, 445-450.
3. Bormuth, John R. "Close Test Readability: Criterion Reference Scores," ***Journal of Educational Measurement,*** 5 (Fall 1968), 189-196.
4. Evans, N. Dean. "Individualized Reading—Myths and Facts," ***Elementary English,*** 39 (October 1962), 580-583.
5. Fry, Edward B. "A Readability Formula That Saves Time," ***Journal of Reading,*** 11 (April 1968), 513-516.
6. Hunt, Lyman C. "The Individualized Reading Program: A Perspective," in L. C. Hunt (Ed.), ***The Individualized Reading Program: A Guide for Classroom Teaching,*** 1966 Proceedings, Volume 11, Part 3. Newark, Delaware: International Reading Association, 1967, 1-6.
7. Jacobs, Leland. "Individualized Reading is Not a Thing," in Alice Miel (Ed.), ***Individualizing Reading Practices.*** New York: Teachers College Press, 1958, 1-17.
8. Lazar, May, Marcella Draper, and Louise Schwietert. ***A Practical Guide to Individualized Reading.*** New York: Board of Education of the City of New York, 1960.
9. McCracken, Robert A. ***The Standard Reading Inventory.*** Klamath Falls, Oregon: Klamath Printing, 1966.
10. McCracken, Robert A. "Basic Principles of Reading Instruction in the Seventh Grade," ***High Trails: Teacher's Edition,*** Sheldon Basic Reading Series. Boston: Allyn and Bacon, 1968.
11. Miel, Alice (Ed.). ***Individualizing Reading Practices.*** New York: Teachers College Press, Columbia University, 1958.
12. Olson, Willard C. ***Child Development.*** Boston: Heath, 1949.
13. Olson, Willard C. "Seeking, Self-selection, and Pacing in the Use of Books by Children," ***Packet.*** Boston: Heath, 1952, 3-10.
14. Sheldon, William D. et al. ***Sheldon Basic Reading Series.*** Boston: Allyn and Bacon, 1965.
15. Spache, George. ***Spache Diagnostic Reading Scales.*** Monterey, California: California Test Bureau, 1964.
16. Tinney, Ronald. "A Study Comparing Instructional Reading Levels With Difficulty Levels of Materials Children are Expected to Read," unpublished paper, University of Victoria, 1970.
17. Veatch, Jeannette. ***Individualizing Your Reading Program.*** New York: G. P. Putnam's Sons, 1959.
18. Veatch, Jeannette, and Phillip Acinapuro. ***Reading in the Elementary School.*** New York: Ronald Press, 1966.
19. Vite, Irene. "A Primary Teacher's Experience," in Alice Miel (Ed.), ***Individualizing Reading Practices.*** New York: Teachers College Press, 1958, 18-43.
20. West, Roland. ***Individualized Reading Instruction.*** Port Washington, New York: Kennikat Press, 1964.

Martin H. Jason
and
Beatrice Dubnow
Roosevelt University

THE RELATIONSHIP BETWEEN SELF-PERCEPTIONS OF READING ABILITIES AND READING ACHIEVEMENT

The self-concept, as operationally defined in various studies, has received the primary focus of research efforts in the area of self-perception and reading achievement. Treated as a global variable reflecting a pupil's generalized view of himself, the self-concept has been reported as being positively related to reading achievement in the majority of investigations. Studies in which the relationship has been supported at various grade levels include Bodwin (*2*), Lumpkin (*8*), Lamy (*7*), Wattenberg and Clifford (*10*), and Williams and Cole (*12*).

A review of the literature has not disclosed any studies employing self-report scales which involve specific reading abilities. One projective instrument (Reading Apperception Test) that dealt specifically with reading was developed by Hake (*5*). The test, designed to evaluate covert motivations of good and poor readers, contains ten ambiguously drawn pictures depicting children in various reading situations. When the instrument was administered to a sample of 80 sixth grade pupils, the results revealed, among other findings, that below average readers had significantly lower self-concepts than above average readers.

Since no self-report scale involving reading appeared to be available, it was the basic intent of the present investigation to develop such an instrument and begin initial testing in order to make judgments concerning its potential usefulness.

The theoretical base underlying the present investigation is derived primarily from the phenomenological principle that the

phenomenal self, as the organization of all perceptions an individual has about himself in a particular situation, governs his behavior in that situation (*3*). What is relevant in terms of understanding inadequate reading performance is that while a pupil's difficulties may be a function of perceptions commensurate with that performance, these perceptions, notwithstanding, satisfy a basic need. What would then on its surface appear as self-defeating pupil behavior is quite the opposite when considered from the phenomenological perspective. Briefly stated, this dimension of the theory as offered by Combs and Snygg holds that since the maintenance and enhancement of the phenomenal self is a fundamental human need, perceptions which are consistent with that self are selected whether they appear complimentary or self-damaging to an outsider. Perceptions which are inconsistent are unlikely to occur as they would not fit the self structure. As applied to the reading situation,

> Most of the cases coming to the reading clinic are poor readers who have nothing whatever wrong with their eyes. They are not unable to read in a physical sense, but are children who for one reason or another have come to ***believe*** they cannot read. What is more, because they see themselves as nonreaders, they approach reading expecting to do badly, and a fine vicious circle gets established . . . .

This cyclical effect is also indicated by Quandt (*9*) who states that "Children . . . who come to school believing that they will not succeed in reading, as well as children who gain this concept at a later time, may become victims of a self-fulfilling prophecy. Believing that they will not succeed in reading their behaviors and efforts during reading instruction contribute to making their expectations come true." In this same regard, "A child who, for whatever reason, develops negative self-perceptions may see himself as an inadequate reader, incapable of learning, or just generally inadequate" (*1*). More positively, "If the child is highly proficient in extracting ideas from the printed page and he recognizes this, he will have a positive approach to reading. He is able to read, therefore his concept of himself is as a 'reader' " (*6*).

The application of phenomenological theory is reflected in the assumption that the perceptions a pupil holds regarding his reading abilities serve to either facilitate or inhibit his reading performance. The following research hypotheses were formulated to test

this assumption: There is a positive relationship between self-perceptions of reading abilities and achievement in 1) vocabulary and 2) reading comprehension. The same prediction was made separately for boys and girls.

METHOD

The Self-Report Reading Scale, a 20-item instrument requiring "Yes" or "No" responses, was designed for group administration. Its purpose was to measure elementary school pupils' perceptions of their reading abilities. Representative items include:

> Most of the time I feel I will never be a good reader in school.
>
> I feel that there are too many hard words for me to learn in the stories I read.
>
> I can read as fast as the good readers.
>
> Most of the time when I see a new word I can sound it out by myself.

The pupil was given one point for each item to which he gave an answer representing a positive self-perception. In order to help insure that a pupil's perceptions would not be inaccurately reported because of difficulty with vocabulary, words above a third grade reading level were not included. The split-half reliability of the Self-Report Reading Scale corrected for test length was 0.88 for the group of fifth graders participating in the study. Other instruments imployed in the study included the Otis-Lennon Mental Ability Test, Elementary II Level, Form J, and the Iowa Tests of Basic Skills, Vocabulary and Reading Comprehension tests, Form 3.

All nine fifth grade classes from a Chicago suburban school district participated in the study. The pupils were grouped according to a multi-age plan. Of the 247 pupils in these classes, 231 were present for all of the testing and only their scores were analyzed.

Arrangements were made to have all testing done with only fifth graders present. Teachers in each of the nine classes administered the achievement and IQ tests during the week prior to the administration of the Self-Report Reading Scale. One of the investigators administered this instrument, which took approximately 15 minutes to complete.

RESULTS

The correlations between scores on the Self-Report Reading Scale and on the reading achievement tests are shown in Table 1. Both zero-order correlations and partial correlations with IQ partialed out are given.

TABLE 1

Means, SDs, and Intercorrelations of Self Report Reading Scale and Reading Achievement

Test	Mean	SD	Correlation with Self Report			
			IQ Not Controlled		IQ Controlled	
			Boys (N = 114)			
Self Report	12.52	4.12	r	p	r	p
Vocabulary	21.11	7.61	.36	.001	.19	.02
Comprehension	32.59	13.20	.34	.001	.15	.06
			Girls (N = 117)			
Self Report	13.27	5.17	r	p	r	p
Vocabulary	22.73	6.94	.58	.001	.37	.001
Comprehension	37.41	12.43	.52	.001	.28	.002
			Both Boys and Girls (N = 231)			
Self Report	12.90	4.68	r	p	r	p
Vocabulary	21.93	7.30	.48	.001	.28	.001
Comprehension	35.03	13.01	.44	.001	.22	.001

An examination of Table 1 reveals significant, although not high, relationships between self-perceptions of reading abilities and "vocabulary" and "comprehension." When IQ was partialled out, the relationships were still significant at the .02 level or less except where "boys' vocabulary" revealed a .06 level of probability. An analysis of correlations of the Self-Report Reading Scale with "vocabulary" indicated that they were not significantly different from corresponding correlations with "comprehension."

CONCLUSIONS

The results to a certain degree support the hypotheses which predicted that there is a positive relationship between self-report measures concerning reading abilities and reading achievement.

Although the lower coefficients obtained after IQ was partialled out permit only tentative conclusions, the findings do indicate a consistent trend in the predicted direction.

The fact that coefficients were higher in all analyses involving girls may be related to the overall superiority of girls in reading achievement which is evident beyond first grade and continues throughout the elementary grades (*4*, *11*).

In terms of further research, experimental efforts may reveal the extent to which negative perceptions could be changed. An analysis of gains made from pretest to posttest in self-perceptions of reading abilities and achievement would yield additional data on the relationship between these variables. Concomitant improvement in both areas could indicate information on the role of self-perceptions as an intervening variable, i.e., one that would have the pivotal effect of influencing achievement positively.

In conclusion it is felt that the Self-Report Reading Scale could be useful in sensitizing teachers to the importance of self-perceptions in the reading process. By becoming apprised of perceptions pupils hold, teachers could utilize the information from this instrument in remedial or individualized programs. In this regard an examination of responses for individual items might provide further direction for the diagnostic process. Through the child's identification of certain areas of concern to him, tests which diagnose specific areas of deficiency in greater depth can next be employed. The instrument may thus serve its best purpose if it facilitates the communication of poor reader's feelings of inadequacy to his teacher.

References

1. Beretta, Shirley. "Self-Concept Development in the Reading Program," ***Reading Teacher***, 23 (December 1970), 232-239.
2. Bodwin, Raymond F. "The Relationship Between Immature Self-Concept and Certain Educational Disabilities," unpublished doctoral dissertation, Michigan State University, 1957.
3. Combs, Arthur W., and Donald Snygg. ***Individual Behavior.*** New York: Harper & Brothers, 1959.
4. Gates, Arthur I. "Sex Differences in Reading Ability," ***Elementary School Journal,*** 61 (May 1961), 431-434.
5. Hake, James M. "Covert Motivations of Good and Poor Readers," ***Reading Teacher,*** 22 (May 1969), 731-738.
6. Homze, Alma Cross. "Reading and the Self-Concept," ***Elementary English,*** 39 (March 1962), 210-215.

7. Lamy, Mary W. "Relationship of Self-Perceptions of Early Primary Children to Achievement in Reading," unpublished doctoral dissertation, University of Florida, 1962.
8. Lumpkin, Donavon D. "Relationship of Self-Concept to Achievement in Reading," unpublished doctoral dissertation, University of Southern California, 1959.
9. Quandt, Ivan. ***Self-Concept and Reading.*** Newark, Delaware: International Reading Association, 1972, 9.
10. Wattenberg, William W., and Clare Clifford. "Relation of Self-Concepts to Beginning Achievement in Reading," ***Child Development,*** 35 (June 1964), 461-467.
11. Weintraub, Samuel. "What Research Says to the Reading Teacher," ***Reading Teacher,*** 19 (November 1966), 155-165.
12. Williams, Robert L., and Spurgeon Cole. "Self-Concept and School Adjustment," ***Personnel and Guidance Journal,*** 46 (January 1968), 478-481.