DOCUMENT RESUME

ED 073 433 CS 000 391

AUTHOR	Farr, Roger; Anastasiow, Nicholas
TITLE	Tests of Reading Readiness and Achievement: A Review and Evaluation. Reading Aids Series.
INSTITUTION	International Reading Association, Newark, Del.
PUB DATE	69
NOTE	54p.
AVAILABLE FROM	International Reading Association, 6 Tyre Avenue, Newark, Del. 19711 (Order No. 208, $2.00 non-member, $1.75 member)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Achievement Tests; *Evaluation Criteria; *Reading Readiness Tests; *Reading Tests; *Test Reviews; Test Selection

ABSTRACT

This book is intended primarily for classroom teachers and other personnel who work directly with teachers in selecting reading readiness tests or achievement tests. The first chapter lists and briefly explains the criteria used by the authors in reviewing the tests included. These criteria are concerned with norms, standardization, objectivity, ease of administration and scorability, validity, reliability, and the test manual. The reading readiness tests reviewed are the Gates-MacGinitie Readiness Skills Test, the Harrison-Stroud Reading Readiness Profile, the Lee-Clark Reading Readiness Test, the Metropolitan Readiness Test, and the Murphy-Durrell Reading Readiness Analysis. The following reading achievement tests are reviewed: the California Reading Tests, the Gates-MacGinitie Reading Tests, the Iowa Silent Reading Test, the Metropolitan Achievement Tests--Reading, and the Stanford Achievement Tests--Reading. The appendix contains two charts, one a general description of the tests reviewed and the other a summary of the technical evaluation of the tests. (TO)

Reading Aids Series

VERNON L. SIMULA, *Editor*

Tests of Reading Readiness and Achievement:

A Review and Evaluation

by Roger Farr and Nicholas Anastasiow

Institute for Child Study
Indiana University

Test Review Contributions by

Charles Mangrum
Miami University

Diane Seltzer
University of Nebraska

An IRA Service Bulletin

Published by the

INTERNATIONAL READING ASSOCIATION • Newark, Delaware

Printed in the United States of America
Second printing, February 1971

FOREWORD

Teachers, supervisors, and administrators are often faced with the task of selecting and administering tests and interpreting their results. That tests do perform a useful function is indicated by their widespread use. Nevertheless, the user is faced with a number of frustrations in selecting and interpreting tests relating most often to his relative lack of background in tests and measurement. Evaluation is the key to the practical value of this publication. The authors present more than a simple review of tests and their manuals. They react critically to what is presented and to what is omitted. They point out the limitations of specific tests as well as the strengths. The person who has serenely accepted tests and their manuals at face value will be surprised and at times perhaps even shocked by what he reads. The objective of providing a useful aid for the reading teacher has been well met with this publication.

The International Reading Association is also publishing an extensive evaluation and review of the research on tests and measurement in reading which will appear in the ERIC/CRIER *Reading Review Series.* This volume is also authored by Roger Farr. Two other titles in IRA's *Reading Aids Series* relate to evaluation in reading and may be of interest to the readers of this bulletin: *Informal Reading Inventories,* by Marjorie Johnson and Roy Kress, and *Evaluating Reading and Study Skills in the Secondary Classroom,* by Ruth Viox.

Leo Fay, *President*
International Reading Association
1968-1969

CONTENTS

Chapter 1

CRITERIA FOR REVIEWING TESTS

• Why Such A Book As This One?

THIS book is intended primarily for classroom teachers and other personnel who work directly with teachers in selecting reading readiness or achievement tests. One may ask, does one really need a guide to select a test? All readers have probably had a course in tests and measurements and know the general rules for selecting an achievement test. However, many had the course before actually teaching so that theory was too removed from practice and therefore, was not so useful as it could have been. But, more importantly, test development has made rapid advancement in theory and practice in recent years.

Selecting a reading readiness or achievement test is continually becoming a more complex task with these advancements. Test manufacturing has become a large scale enterprise with attractive and highly promoted reading achievement, assessment, and diagnostic devices. Some of these instruments are based on new research evidence on how children learn to read. Other tests are designed specifically to measure experimental programs, rather than the more traditional approaches.

The computer has also made an impact on test construction. Rapid analyses of the statistical characteristics of a test are now possible. In the past it would have taken months or years to analyze the results of each item on a test given to a large sample of children. Using rapid analysis techniques, the computer has enabled test manufacturers to revise their tests more frequently, and the revision of old tests is based on more accurate and complete information about the effectiveness of each test question.

Old tests, however, remain in the schools long after the curriculum has changed. These tests are outdated and no longer serve the purpose for which they were originally designed. Yet, on the other hand, some of the older tests still are the "best" that are currently available. How does a teacher choose among them? Selecting a test takes time and careful evaluation, more time than the classroom teacher has to give from his other instructional duties. This book is designed to review the major issues that should be considered before a test is chosen as the one to be used in a classroom.

The authors have reviewed several of the most commonly used reading readiness and achievement tests currently available and have evaluated these instruments from both their content and statistical characteristics. An analysis of the research reports from the ERIC Clearinghouse on Retrieval of Information and Evaluation on Reading was used in an attempt to determine which reading tests were being used most often.

These test reviews will hopefully serve as a guide for evaluation in selecting the appropriate test for use in a specific classroom. This guide should reduce the time normally spent in evaluating a reading readiness or achievement test. The issues considered by the reviewers in evaluating the tests are the content measured by the test, its statistical properties, its scorability, the meaning of the subtest and total test scores, and whether the test measures adequately what it purports to measure. Although the results are summarized, it may be useful to review the purposes and uses of achievement tests.

• Why Use a Commercially Prepared Reading Achievement Test?

Prediction and Assessment

One's observation of a child's daily performance is the main source for determining how well a child is doing. However, one will also want to make periodic controlled assessment of each child's current reading ability in order to place him at his appropriate instructional level. Teachers are aware that a child makes the most rapid progress when instruction is near his current level of mastery. Thus, tests help teachers make initial, rough assessments so that instruction can begin with a better probability of success.

Teacher-made tests are one of the main sources of gathering data about children in a classroom. These results help one to predict future achievement, assess how well children have accomplished the goals, provide feedback to the child, as well as reinforce the student for what he has accomplished. However useful these results may be, teachers, parents, and administrators are prone to want some outside assessment of how well the students are doing when compared with a large sample of children of the same age and grade. Teachers have available a limited number of children in a class to compare how well that class or an individual student is progressing. Thus, commercially prepared tests are used to provide wider prediction and assessment of the pupils in a class.

There are other uses of tests besides those listed previously. A school district may wish to look at the general achievement level of its students. This district assessment may help the administration make suggestions for program improvement, purchasing additional instructional aids and equipment, or providing additional personnel. In addition, tests are used for research purposes to evaluate the effectiveness of a new program or to compare two modes of instruction. Any of the criteria to be described are relevant for these uses of tests as well.

• Factors to be Considered in Choosing a Test

Norms

A commercially prepared test usually offers the advantage of having been administered to a large number of children from a wide variety of rural and urban centers. Usually these tests have been administered to children of various social, racial, and ability levels. Thus, the test will have been "normed" on a population of children from more than just one class, school district, or state. A description of the norming population is critical for an interpretation of test scores. If one has a bright, urban class and the test has been normed with average, intercity children, the scores one's children obtain may indicate higher grade scores for that class than is a realistic assessment. If the reverse is true, that the test originally has been given to a large population of bright youngsters, the scores may be lower than is a realistic appraisal of one's students' current status.

Standardization

Adequacy of prediction and assessment are pertinent considerations for selecting a test. Often this category is called standardization, a term which is not an accurate description of one very important aspect of the test that one is concerned with.

Clear standardized directions on how the test is to be administered are desirable. A set of directions that is concise and uniform will ensure that the results are not depressed or inflated because the directions left the procedure unclear. The students' scores will not be so useful if the test is given in different way from the way it was given to the norming population.

Objectivity

A commercially prepared test also is intended to be objective; i.e., the score achieved should not be biased in some way by the tester or observer of the child's demonstration of what he knows. Encouragement, as everyone knows, can guide a pupil to a right answer. This is an excellent instructional technique as guided-discovery experiments have demonstrated. However, at times one will want to know not how much a pupil can learn but how much he has learned and where he is now. An objective measure should enable one to determine this. As one shall see, tests vary in their objectivity.

Ease of Administration and Scorability

Given enough time and personnel, a teacher might collect very extensive data about a child. This undertaking is not possible in most instances. Teachers want a test that makes reasonable demands in terms of the amount of time needed to administer the test so that children are not fatigued and also so the classroom instructional program may continue. In addition, tests that are difficult and tedious to score are sources of error

and use far more teacher time than is desirable. Most achievement tests are designed to minimize the scoring time required of teachers.

Validity

The test selected should measure the content one is teaching. A reading readiness test should predict success in initial reading–not in group co-operative play, although the two behaviors may be related. In addition, the test should predict success in the classroom, whether one uses a look-say approach or a more linguistically oriented program.

A reading achievement test should sample the decoding, vocabulary, and comprehension skills taught. The titles of the tests should be an accurate description of the skills being tested. Proof should be given that the skills of the test were measured with the norming population.

The test should provide evidence that the skills measured are either a measure of current status or are of predictive value. One should know which tests can be used to predict success or failure in subsequent instruction. Not all tests provide this kind of evidence.

Three kinds of "validity" are important to consider. One is content validity, which assesses whether the test measures the content one is teaching. Second is concurrent validity which compares the test behavior to current performance. The third is predictive validity, which tells whether the score the child receives can be used to predict how well he will do in the future. A fourth, more difficult kind of validity, is construct validity, which refers to the psychological processes represented by the behaviors exhibited by the child during the test. For example, some reading tests claim that the comprehension skills measured on the test evaluate the child's ability to make inferences. Evidence should be offered by the test manufacturer that the questions on the test do measure this trait.

Reliability

When choosing a test one will want it to be a reliable measure of how much a child knows or how well he is able to apply his skills. The test results should not be a chance score with a child obtaining a high score by luck, guessing, or other factors. The test should not be constructed so that it gives the advantage to children who know only one thing well. The test should be constructed so that one has confidence that the score the child receives today will be similar to the score he would receive if the test were to be readministered to the same child tomorrow.

The Test Manual

It is the professional responsibility of the test maker to provide sufficient and appropriate evidence for the user to judge whether a test fits his purposes. Description of administration, norming, scoring, reliability, and validity should be provided in the user's manual. The authors have used the test manuals to evaluate the evidence provided and to assess in what ways the test can be recommended for use.

• How Can One Use Test Results?

Most achievement tests are group tests and provide a rough assessment of how a child compares with the norming sample. Such tests are not meant to be diagnostic, nor are they meant to give an accurate assessment of functional reading levels. They are a rough and ready means of grouping children for reading instruction. The grade placement score has little instructional value. The percentile score is more useful but again requires careful interpretation. If a test is used over a period of time, class norms may be built for a particular school district.

One of the greater *misuses* of the group standardized reading tests is the use of grade level norms as an indication of the level at which a student ought to be given reading instruction. Because of the nature of standardized tests, they are not appropriate for determining the reading level at which the youngster can profitably receive instruction. Standardized reading tests are developed from a group of items which are administered to a particular norming group; the grade norm is based on the average number of items that students get correct at a particular grade level. For example a score of 6.0 would only indicate that a youngster who is just beginning sixth grade had 100 items correct. This score does not mean that the student who had 100 items correct can necessarily read 6.0 grade level material. The standardized tests were not meant to be criterion tests!!

What we are suggesting is a procedure that might be used to determine the level at which a youngster may be given instruction on the basis of his standardized reading test score. Betts, in his 1942 book *Foundations of Reading Instruction,* suggested three functional reading levels. These functional reading levels are based on work that he and Patrick Killgallon had done. Credit also is given to Thorndike for the idea.

The three functional reading levels are as follows: 1) The *independent* reading level, the level at which a youngster should be doing his leisure-time reading; 2) *instructional* reading level, the level at which the youngster should be given reading instruction and should be learning in the content areas; 3) the *frustration level,* the reading level which is too difficult for the youngster and which will probably lead to negative conditioning to reading.

The *independent level* is identified by 99 percent or better word call, 90 percent or better comprehension, and freedom from behavioral symptoms of tension and anxiety. The *instructional level* is identified by 95 percent or better correct word call, 75 percent or better comprehension, and only slight signs of anxiety. The *frustration level* implies 90 percent or less correct word call, less than 75 percent comprehension, and symptoms of nervousness, anxiety, and frustration.

A grade level score from a standardized reading test more often than not places a youngster at his frustration reading level. This relationship, of course, is dependent on the particular standardized test that is used and the particular material which is used for the informal reading inventory.

A procedure which might be used by classroom teachers to *determine the functional reading levels* (i.e., independent, instructional, and frustration) that correspond to *various scores on the standardized tests* would work something like this:

The teacher would administer the usual standardized test to his class. He would then administer an informal reading inventory *(IRI)* to some of his students: the informal reading inventory should preferably be based on the basal reader which he was using for instruction. Youngsters to be tested with the *IRI* would be selected from several points along the range of scores students achieved on the standardized tests. Students should be selected for testing at least from the bottom, middle, and top of the range of scores. Additional points on the range could be sampled if time allowed. The teacher would then determine the relationship between *various raw scores* on the standardized reading tests and the *functional reading levels on the informal reading inventory*. After he has gathered data of this sort for several classes, he would not find it necessary to readminister the informal reading inventory but could use the past performance of students to determine the levels at which they ought to be given instruction.

These procedures would result in the teacher's ability to determine a student's functional reading level that would correspond to a particular raw score on a particular standardized reading test. For example, a student who scores 121 raw score points on a standardized reading test might have a fourth grade independent reading level, a fifth grade instructional reading level, and a sixth grade frustration level. Such knowledge would allow the teacher to utilize the standardized test scores to place each student at the instructional reading level where he would have the greatest opportunity to succeed.

• Plan of this Reading Aid

Each test included in this review was assessed using the following outline:

I. Test overview
   - A. Title
   - B. Author(s)
   - C. Publisher
   - D. Date of publication—original, revised
     - 1. Manual
     - 2. Test
   - E. Level and Forms
     - 1. Grade level
     - 2. Individual or group
     - 3. Number of forms available
   - F. Administration Time
   - G. Scoring—hand or machine scorable
   - H. Cost
     - 1. Question booklets—consumable or not

H. Cost (cont.)
   2. Answer sheets
   3. Manual

II. Evaluation of Subtests and Items
   A. Description of subtests
      1. Given meaningful name—describe test adequately
      2. Is each subtest long enough to provide usable results?
      3. Sequential development of each subtest logical, and transitions smooth?
   B. Author's purpose reflected in selection of items
   C. Scoring ease and usability of tables
   D. Directions—clarity and level of language appropriate to grade level
   E. Design—format, currentness, printing, legibility, pictures
   F. Readability

III. Evaluation of Reliability and Validity
   A. Norming population
      1. Size
      2. Age, grade, sex
      3. Range of ability
      4. Socioeconomic level
      5. Date of administration
   B. Validity
      1. Content validity
         a. Face validity
         b. Logical or sampling validity
      2. Empirical validity
         a. Concurrent
         b. Predictive
      3. Construct validity
         a. Construct and theory of which construct is a part clearly defined
         b. Discriminant or convergent validity evidence
         c. Significant difference found in performance between groups which have varying degrees of this trait?
      4. Does reported validity appear adequate in relation to author's stated purpose? Why or why not?

Following this review, tables (refer to Appendix) were constructed to summarize the characteristics of the test for a quick-and-ready reference for use.

Each test is described, and the strengths and weaknesses are delineated so that one may evaluate the test one's self. Each review was sent to the publisher for his reactions. In some cases, additional information was given the authors and this matter was included in the review. If the necessary data were not located in the manual but found elsewhere, the appropriate sources have been indicated. If the authors did not agree with the publishers' criticisms, this fact has been indicated so that potential test

users can come to their own conclusions. It should be the teacher who makes the final decision on the use of a test based on his program; the authors can only guide and suggest the criteria by which that decision might be made.

• What is the Responsibility of Test Publishers?

A test should be placed in the same category as a critical drug. A test should be used only after a careful study of its effects has been made. Evidence should be provided that the test (or drug) will do what it purports to do. Too many critical decisions are made about a child based on his test scores to use any test that is not a reliable and valid assessment of the child's ability to do the task described by the test. A teacher should insist that the test manufacturers provide him with the same reputable product that he would demand of a drug manufacturer who offers a new cure. It is better to use no test than to use an unreliable or invalid one. One finds that a number of tests are released before adequate data are available.

Many tests have not been studied sufficiently before they are put on the market for sale. One hopes the reader will note these deficiencies and realize how serious the action is to make an instructional, promotional, or evaluational decision about a child when it is not based on an accurate, stable, or predictive measure of his achievement.

Chapter 2

SELECTING READING READINESS TESTS

READING readiness tests are administered in a majority of the elementary schools in the United States. The popularity of the tests alone suggests that they are useful to teachers in making decisions about children. What does the teacher who administers these tests hope to learn about the students? First, he would like to know if the score on the test is a valid predictor of whether a particular student is ready to begin formal reading instruction. In addition, the subtest scores on the readiness test are said to assist in diagnosing the readiness skills in which each student is weak or strong so appropriate instruction can be planned.

These two reasons should, therefore, be the prime considerations in evaluating a reading readiness test. One should seek evidence that relates to the predictive power of the test. Do students who score well on the test become good readers? Are these high scorers ready for formal reading instruction? Secondly, one should examine the subtests and items to determine if one agrees that these are the most important skills for a student to develop if initial reading instruction is to be successful. If it is decided that the skills from the test are appropriate, then one should look for evidence regarding the uniqueness of the subtests. In order to use the test in a diagnostic fashion, the publisher should provide evidence that the subtests deal with separate measurable skills.

Further evaluation of a reading readiness test should include a more careful than usual examination of the testing procedures and the test format. Because reading readiness tests are used with such a young age group, the examinees can easily be penalized by an unusual test format or a lack of clarity in the examiner's testing procedures.

Finally, one should also examine the usefulness of the test scores. This aspect is partly determined by the subtests included on the test, but it is also determined by the description of the use of the test provided by the publisher. One should feel confident in knowing what to do with the test scores. How do they relate to reading readiness? Do low scores mean that a student should not begin reading instruction? How should the subtest scores be used? All of these questions should be answered by the publisher in a clear statement. In addition, the publisher should discuss the relationship of the readiness skills measured by his test to the readiness skills or child behaviors which cannot be measured by a test. Lack of such a discussion will seriously limit the use of the test.

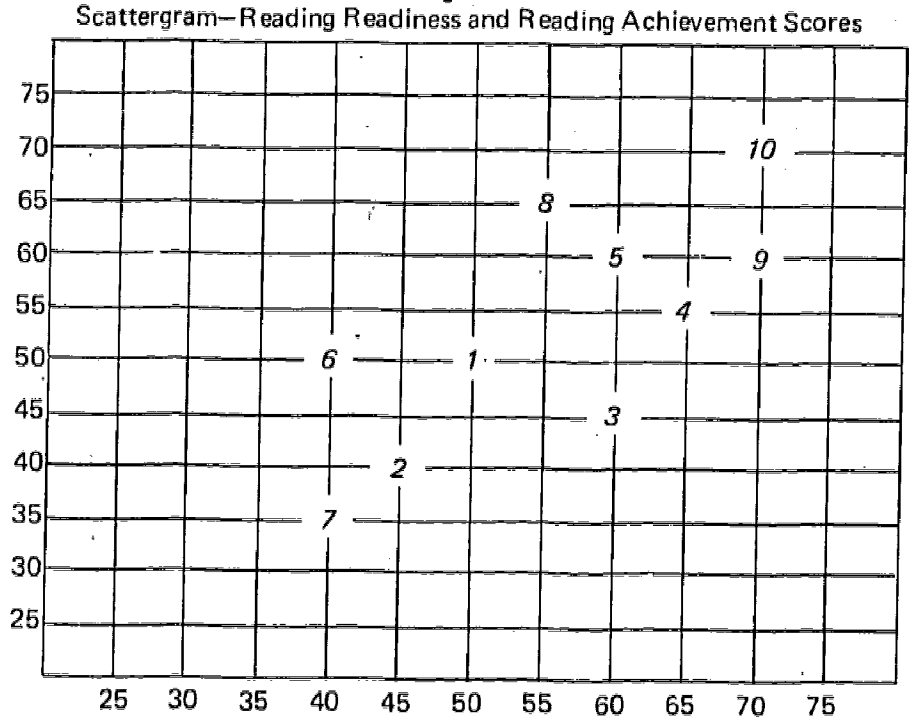
It is believed that the most important use of reading readiness tests is to determine which readiness skills need further development before the students can begin to learn to read. However, the predictive validity of a readiness test should help one to determine the importance of various readiness skills. Most publishers of reading readiness tests do describe such validity.

Regardless of how much evidence is provided, a simple check on how well the test predicts for a class can be obtained. After administering the scores and beginning instruction, one should save the tests until the end of the year and then compare the reading level obtained at the end of the year with the readiness score. Is it in general agreement with a class's achievement? Are there some students who failed to do as well as predicted? Are there some who did better? Can the reason for this be determined?

The use of a scattergram (Figure 1) will help one to visualize how well a particular reading readiness test predicts reading achievement for students in a specific instructional program. The scattergram is developed by plotting the child's intersection of his readiness test score with his scores on a subsequently administered reading achievement test. The readiness test may have been administered at the beginning of first grade and the reading achievement test at the end of first grade, but this time is dependent on the period for which one would like to predict. The scattergram will be more useful if one uses a number to represent each student

Figure 1

Scattergram—Reading Readiness and Reading Achievement Scores



rather than an "X" for all students. If the test publisher provides standard scores or percentiles it would be better to use these rather than raw scores in plotting the scattergram.

The scattergram includes ten students, each represented by a different number; the scores indicate that the readiness test does predict reading achievement scores fairly well. For example, child number seven scored somewhat poorly on the readiness test and on the achievement test; child number ten scored fairly high on both tests.

The charts in the Appendix describe and evaluate several of the major aspects of the readiness tests reviewed. A more complete evaluation of each test follows.

• Gates-MacGinitie Readiness Skills Test

Overview

The Gates-MacGinitie Readiness Skills Test is a revision of the Gates Reading Readiness Test. The new test, published in 1968, is intended for use with pupils at the end of kindergarten or the beginning of first grade. Eight subtests are included, but only the first seven are combined to arrive at a total readiness score. The seven required subtests are *Listening Comprehension, Auditory Discrimination, Visual Discrimination, Following Directions, Letter Recognition, Visual-Motor Coordination,* and *Word Recognition.*

Student responses are recorded in the test booklet. The pictures and words are large and easy to read; however, the use of some aid to help a student keep his place would probably aid in the administration of the test. The directions for the examiner and the oral directions to the examinees are adequate in terms of clarity, completeness, and appropriateness for kindergarten and first grade students. A separate scoring key is provided, and tables are included for end of kindergarten and beginning of first grade; however, the norm group is not described for either of these populations. The tables provide stanine scores for the subtests and total score as well as a percentile score for the total score.

The total raw score is arrived at by multiplying each of the subtest scores by a weighting factor of from one to three. This scoring procedure is used because the test authors feel that certain subtests are more predictive of later reading achievement than others. The *Letter Recognition* score, for example, is multiplied by three while *Listening Comprehension* score is multiplied by only one. This procedure was developed by analyzing data from the standardization of the test. However, no information about this study is reported in the manual, and the reading achievement test which was used as the criteria test was not named. Because of this limitation, it is difficult to determine the value of the weighting procedure. If one utilizes this test, it would be useful to compare the total raw score with the weighted total score to determine which is the better predictor of later reading achievement with one's classes.

Norms

While norms are provided for both the end of kindergarten and the beginning of first grade, there is no descriptive information of the norming population. There is not even an indication of the total number of cases included in the sample. There is also no reference made to any technical information available from the publisher. For these reasons, it would be very unwise to use the test norms provided by the publisher. One would be comparing one's students to a completely unknown population, and this knowledge will not be of any value in determining whether these students are ready to begin reading instruction.

Validity

The content validity of the test appears to be appropriate for measuring many of the skills necessary to beginning formal instruction in reading. Several of the pictures seem to be biased toward a middle-class population. Ethnic differences are represented with several pictures of Negro children. The results of using the test with certain cultural groups would aid in determining the validity of the test with these special groups. The authors also encourage the use of teacher observations and informal tests for measuring other aspects of the pupils' development. For these reasons, the content validity of the test is quite satisfactory; however, there is a complete lack of any other validity evidence. This condition would make any diagnostic or predictive use one might make of the tests completely dependent on the information gathered with one's own classes.

Reliability

The authors discuss some of the pertinent factors related to reliability—such as, the higher reliability when test scores are combined rather than used separately, the higher reliability with relatively longer tests, the higher reliability of scores in the middle of a range of scores when compared to scores at either extreme, and the relatively high unreliability of differences between test scores. This information is well presented and should be considered by the test user: but the publisher does not give any information about the reliability of the total test or subtests, and, therefore, there is no basis on which to determine if the score a pupil receives on one day is likely to be the same as the score he would receive on another day.

Evaluation of Subtests and Items

The names of the subtests adequately describe the tasks. Each test is arranged in a logical order, and generally the tasks become more difficult as each test progresses. The use of letters and words in the *Visual-Discrimination* and *Visual-Motor Coordination* subtests seems to be in keeping with the trend away from the use of geometric shapes as was the case with earlier readiness tests. This procedure will probably increase the predictive validity of the tests because the tasks more closely resemble actual reading behavior.

The authors caution against the use of separate subtest scores and suggest that the total test score is more useful. The writer would strongly support this advice because of the lack of reliability and validity evidence for the subtests and also would caution against any diagnostic use of subtest scores, even when stanine scores differ by as much as three stanines as the test authors suggest, because several of the subtests are quite short, varying in length from eighteen to twenty-four items, and also because the norming population on which these stanines are based is a completely unknown quantity. The test authors also encourage the thoughtful interpretation of any student's scores and suggest that reading readiness test scores are quite dependent on the teacher's instructional procedures and everything he knows about the children in his class. It is very refreshing to see such a statement printed in a test manual.

The eighth subtest, which is not used in arriving at a total readiness score, is *Word Recognition*. This test is actually a reading achievement test and can be used by the teacher to identify those children who have already begun to learn to read. While the test is a useful addition, it is quite probable that the alert teacher would not need such a test to identify the student in his class who had already begun to learn to read.

Summary

This test appears to have content validity for measuring many of the skills which are necessary to begin reading instruction. The authors point out the shortcoming of readiness tests and generally do an adequate job of describing the value of the test; but the lack of complete validity, reliability, and norming data make the test of very limited use to the teacher. The test could be used as a criteria test for determining achievement levels for certain readiness skills, but it is probable that the subtests are too short to give valid or reliable information. This test appears to be one that has been published and is available for sale before the collection of validity and reliability data. A more complete test manual would also enhance the value of the test.

• The Harrison-Stroud Reading Readiness Profile

Overview

The Harrison-Stroud Reading Readiness Profile is presented in three booklets and was revised for publication in 1956. According to the authors, the test is designed to measure those skills which are necessary for beginning reading. Six subtests are included: the first five can be administered on a group basis, but the sixth must be administered individually. The subtests include *Using Symbols, Making Visual Discriminations, Using the Context, Making Auditory Discriminations, Using Context and Auditory Clues,* and *Giving the Names of the Letters.*

Students' responses are written in the test booklet; only one form of the test is available. Scoring is somewhat difficult because no separate scoring key is provided; the examiner must search through the manual for

the scoring key for each subtest. For convenience, tables for converting raw scores into percentile ranks are printed directly on the front page of the first test booklet. The examiner's directions are clear and precise, and the language of the oral directions is also appropriate for kindergarten and first grade children. The format of the test is attractive and efficient; colored boxes are utilized as place-keeping devices. The use of three colors is functional in giving the directions for each item of the test. The children's layout of the questions is spacious and clear.

Norms

The norms for interpretation of raw scores were based on 1,400 pupils in thirty-two communities in twenty-eight states in 1955. This population is not adequately described as to range of ability, sex, chronological ages, and socioeconomic level. This is a serious weakness of the test and would make the use of the norms very questionable. In addition, the use of 1955 norms also seems to be a dubious basis for evaluating the performance of children today. The authors do not suggest the development of local norms, but this procedure would seem to be essential for interpretation of raw scores because of the cited limitations.

Validity

The content validity of the test is based on the logical assumption that the test measures the skills essential to beginning reading. All of the tests appear to achieve this purpose. The test lacks other validity evidence. It is believed that the most important use of a reading readiness test is to diagnose students' strengths and weaknesses necessary to beginning reading instruction, but the test authors present no evidence regarding the diagnostic or subtest validities of the test. A second use of readiness tests is the prediction of later reading achievement, and this test includes no predictive validity evidence.

Reliability

The complete lack of reliability data is one of the main weaknesses of the manual of this test. The manual gives no evidence regarding the reliability of the total test or of the subtests.

Evaluation of Subtests and Items

The names of the subtests are meaningful and adequately describe the tasks involved. The tasks on the subtests are consistent with the authors' stated concept of the nature of reading readiness. They list eight factors that are important in reading readiness, and all of these factors are included in the subtests. The authors suggest that other factors, such as intelligence test scores and teacher observations, should also be used in determining instructional group placement; but the student behaviors the teacher is to observe are neither described nor discussed. The most serious problem is the lack of a total readiness score. The authors defend this procedure by suggesting that the tested skills do not develop evenly in

children and, therefore, suggest that the subtests should be used diagnostically. However, the use of subtest scores in this manner would mandate evidence of the distinctness of each subskill and also evidence regarding the reliability of each subtest, but none is given. The authors also suggest the use of percentile levels for grouping pupils for instruction based on performance on the subtests, but they equate each test as being of equal importance in this procedure.

The *Using Symbols* and *Visual Discrimination* subtests utilize words rather than geometric shapes as many readiness tests do. It is believed that using words is the more valid procedure because of its more closely resembling actual reading behavior. Most of the items on the test do not appear to be extremely biased toward a middle-class culture; however, evidence regarding this conclusion is not available. Some of the items are a bit dated but do not appear as though they would interfere with the pupil's determination of the correct responses.

The lack of reliability for the subtests is compounded by the authors' suggestion that the subtests be used diagnostically and also by the relatively short length of most of the subtests. The total number of possible correct items for each subtest is 22, 14, 16, 18, 16, and 18.

Summary

The test does have face validity for the diagnosis of reading readiness skills. The lack of reliability and validity evidence seriously limits the value of the test. The use of the test norms is not recommended because of the limited description. The writer suggests that this test could be used most effectively as a criteria test for measuring mastery of certain skills, but it should not be used for comparative purposes unless local norms are developed for that purpose by the teacher or school district.

• Lee-Clark Reading Readiness Test

Overview

This is one of the better known readiness tests. Its reputation is largely due to the many editions that have been available since as early as 1931. It is the 1962 edition that this review covers.

The test is composed of three parts made up of four subtests. Part I contains two 12-item tests of *Letter Symbols,* a total of 24 items. Part II contains a 20-item *Concept* test, and Part III consists of a 20-item *Word Symbol* subtest.

A partial scoring key is provided. For two subscales, scoring is done by inspection without mechanical aids or accessories. The manual suggests that an extra test booklet can be marked or cut out for a scoring stencil. Although this marked or cut out copy would take an examiner time to prepare, it would probably make scoring more convenient. Scores are reported on a profile on the back of each test booklet. The profile provides interpretation of grade placement, expectation of success rating, and indication of months of delay before beginning formal reading instruction.

When completed, the profile provides the teacher with an overall picture of the child's performance on the entire test. Two tables are provided for grade-placement equivalents of total scores and classifications of high or low readiness for entering first grade pupils and for end-of-year kindergarten pupils. The grade placement scale of high, high average, low average, and low was based on a small sample of 177 pupils who were given this readiness test in the first month of first grade. In April and May the sample subjects were given the Lee-Clark Reading Test: Primer. The pupils were then divided into four groups in terms of their primer scores. This sample was quite small, and one must question the lack of further descriptive information pertaining to the sample. One must also be dubious about the validity for such exact and detailed analysis of the four interpretative categories. However, this approach is an improvement over the 1951 edition which provided no statistical support for the interpretative tables.

Directions for administration of the test are clearly and exactly stated in the manual on an appropriate language level for young children. The authors caution examiners to use the exact directions and to administer the tests in small groups. When the group exceeds 15 pupils, the authors recommend an additional adult assistant. The nature of the illustrations has been revised in this 1962 edition to enlarged, shaded drawings. However, the drawings in subtest three are small and blurry. The pictures, done in a soft green shade of ink, produce a pleasing, nonglaring effect. The format is attractive and easy for children to manipulate.

Norms

Norms for this readiness test are based on a "slight adjustment" of the 1951 norms. The 1951 norms were based on 5,000 entering first graders. No further description of the 1951 norms is provided in the revised 1962 edition. Although the 1951 edition provides some information pertaining to the norm sample—such as, median chronological age, median IQ, and racial background—test users should not be required to search out an earlier edition for this important information. Since 1951, the norms have been adjusted to produce slightly more difficult norms on the basis of 1,000 end-of-kindergarten and first grade pupils. Two of the first grade samples were also followed up and tested at the end of the school year for comparison. The slight adjustment of the 1951 norms provides a questionable norm basis for kindergarten and first grade children of 1968. New norms which utilize up-to-date samples would provide a better basis for current norm tables.

Validity

Predictive validity studies are presented for the 1962 revision and for previous editions. The 1962 revision validity studies were based on five groups of entering first grade pupils. However, the manual states that the mean scores for these groups approach the upper limits of the test and that the standard deviations were restricted, such information indicating that

few of the pupils in these classes had scores that placed them below a readiness classification of "high average." Therefore, the reported validity will not be applicable to all test-consumer groups. The manual also reports other predictive validity studies that were conducted, such as the study in the public schools of Portland, Oregon. However, the manual states that ". . . the pupils engaged in this study were *far advanced* in readiness development by the time they were tested." This fact is reflected in the mean Lee Clark Reading Readiness Test scores. This predictive validity study with "far advanced" readiness students does not reflect the recommended test administration time of end-of-kindergarten or entrance-of-first-grade that is stated in the manual. The manual further cautions that ". . . if testing is delayed too far into the first grade . . . the results for many pupils in normal groups will do no more than verify that they are ready to read." The "advanced" readiness sample group will also be inappropriate for many test consumer populations.

The manual does not report any discussion of face validity or sampling validity. To the observer, however, the individual subtests do appear to have appropriate content. According to the defined behavioral terms of the readiness trait in the manual, the subtests do not represent all the aspects of readiness. The test authors suggest that a definition of readiness should include physical maturity, motivation, mental ability, emotional adjustment, and experiential background. Due to the fact that the test measures a more limited number of skills, the authors rightfully recommend that this readiness test should not be the sole measure or basis for decisions on pupils' reading readiness.

Reliability

Reliability coefficients ranging from .87 to .96 were established on the basis of split halves by the Spearman-Brown formula corrected for length. The sample on which these reliability coefficients were based is not described adequately. The following is the only description of the sample: ". . . unselected school samples having means and standard deviations typical of a majority of schools in which the test is administered." The standard error of measurement showed that the chances were two to one that the examinees' readiness grade placements would not vary more than two months, and nineteen to one that they would not vary more than four months from their true readiness grade placement.

Intercorrelation coefficients and reliability coefficients based on the Kuder-Richardson formula were computed for part raw scores. This matter is important for diagnostic use of the part scores. The resulting coefficients showed that Part I *(Letter Symbols)*, Part II *(Concepts)*, and Part III *(Word Symbols)* scores were sufficiently independent for utilization in determining whether pupils have understanding of spoken words and concepts. Considering this test's brevity (it only takes [illegible] minutes to administer), the reliability is surprisingly adequate.

Evaluation of Subtests and Items

Subtest one requires the child to match letters in one column with the next. This test is called *Matching* in the scoring key rather than *Letter Symbols-Part I* as it is named in the manual. The test is short, but the score is combined with test two, a 12-item subtest, which requires the child to match the correct letter out of an array of four letters. The combined 24 items are sufficiently long.

Part II (test three) is named *Concepts* but is referred to as *Cross-out* in the scoring key. The child is presented with a series of pictures and is to "cross out" the item named by the teacher. The quality of the pictures, in the writer's opinion, are somewhat blurred and too detailed for adequate discrimination by young children. It is interesting to note that this subtest has the lowest reliability, too low for individual use; the style of presentation may be a factor. The research on concept learning suggests simple line drawings form the presentation of concepts rather than overly detailed pictures. The pictures are not too dated, but the writer suspects that the inner city or rural, impoverished child would not be familiar with many of the concepts presented. No data are presented that indicate whether the test has been used with a variety of schools. The lack of information on the sample for which these data are based makes it extremely difficult to know for what group to recommend the test. Possibly surburban schools have long used readiness tests, and this is the population on which the reliability was established.

The sequential development of the individual subtests is according to increased difficulty, and the transition between subtests is smooth.

The test authors' stated purpose of this test is not only to predict ability to learn to read but also to provide data for intraclass grouping and to analyze reading readiness needs. It should be noted, however, that the number, nature, and length of the subtests do not lend themselves to giving this information. This readiness test appears to be more useful as a gross screening device rather than as a diagnostic tool. Therefore, the specific items and subtests do not reflect all of the test authors' stated purposes.

Summary

As noted before, it is the test manufacturer's responsibility to provide adequate information about the norming of the test and the nature of the population of students used to establish reliability and validity. From the limited information given, the writer suspects the population consisted of groups of brighter-than-average, middle-class children. Thus, the students' test scores in a class should be evaluated with great care. While the total score reliability is high, the part two concepts score section should not be used alone.

The entire test may be useful as a gross screening device, but it is not sufficiently broad in the skills that it measures to provide diagnostic assessment of children's readiness strengths and weaknesses.

• The Metropolitan Readiness Test

Overview

The Metropolitan Readiness Test is widely known and has recently been revised in 1964. The 1964 edition is reviewed here.

The test gives scores for six subtests and a total score; an optional subtest, *Draw-A-Man* is included. The subtests are *Word Meaning, Listening, Matching, Alphabet, Numbers,* and *Copying.*

A separate scoring key is provided to facilitate hand scoring. A scoring service is also available. No description of the scoring service is provided in the test manual. Tables for converting raw scores into percentile ranks, stanines, and statements of readiness—such as, superior, high normal, average, low normal, and low—are conveniently arranged and are easy to utilize. The directions for administration of the test are clearly and exactly stated in the manual and include directions and illustrations to help score the subjective *Draw-A-Man* subtest. The oral directions to the pupils are on an appropriate language level for young children. The format is adequate for the age level and includes green lines to separate items and symbols to help the children locate and maintain their places in subtests where needed, such as in *Alphabet* and *Numbers.* Although most of the items are current, the majority appear to be drawn from middle-class experiences of suburbia, particularly the eastern part of the United States. In addition, it would appear that some individual items may be measures of intellectual functioning rather than measures of readiness to begin formal reading.

Norms

The norming population consisted of 12,231 pupils in 65 school systems in a wide, regional distribution in the New England, Middle Atlantic, Central, and South Pacific states in 1964. Unlike many test manuals, this one states a caution in the use of the norms because of the slightly higher income median of the sample. The test developers encourage establishment of local interpretative norms based on local experience. This is a very desirable statement as the American Psychological Association's *Standards for Educational and Psychological Tests and Manuals* points out: "Local norms are more important for many uses of tests than are published norms. In such cases the test manual should suggest appropriate emphasis on local norms . . . ."

Validity

In regard to content validity, the term and concept of "readiness" are carefully defined, and a list is provided of the most important components of first grade readiness in the view of the authors. After this breakdown of the total area of readiness into categories, the content of each of the six subtests is discussed against a background of the analysis of the components of readiness. While the content of the scales appears to be

appropriate, the word meaning and listening comprehension scales seem to contain some items that are more suitable for middle-class youngsters.

Evidence of concurrent validity is presented by means of correlations of the Metropolitan subtests and the total score with scores on the Murphy-Durrell Reading Readiness Analysis and the Pintner-Cunningham Primary Mental Ability Test. The total score on the Metropolitan correlated quite highly (.80) with the total score on the Murphy-Durrell. There was also a correlation of .85 between the *Letter Naming* subtest of the Murphy-Durrell and the *Alphabet* subtest of the Metropolitan. Other correlations among the subtests were small and probably indicate that the two readiness tests sample abilities that make up their respective composites in different manners. A correlation of .76 was found between the total score on the Metropolitan and the total score on the Pintner-Cunningham test. Because little information about the sample is given in the manual, the prospective user is not able to judge whether the reported validity is pertinent to his situation. Predictive validity studies are only reported for the three experimental forms rather than for the final forms, A and B. In the experimental forms, the *Alphabet* subtest seemed to be the best predictor of future success in reading. The *Numbers* subtest was a good predictor for both future reading and arithmetic success. Additional predictive validity studies are being conducted now and will be provided in future editions of the manual. These studies are available upon request from the company and will be included in a new manual.

Evaluation of Subtests and Items

The following subtests are included: *Word Meaning, Listening, Matching, Alphabet, Numbers, Copying,* and an optional subtest, *Draw-A-Man.* The names of the subtests are not meaningful. For example, the names *Matching* and *Copying* are not expanded or explicit enough to describe whether words, letters, and/or geometric forms are matched or copied. The *Matching* subtest actually measures visual perception involving the recognition of similarities through the use of words and forms. The *Copying* subtest involves letters, numbers, and forms. The length of each of the subtests is quite short, and the reliability of the individual subtests is much lower than that of the total score. However, the manual *does suitably discourage attachment of significance to the individual subtest scores.* The number and type of subtests appear to be consistent with the purposes of the authors, for they designed this test as a measure of readiness for first grade instruction; therefore, both number and reading readiness factors are considered. The authors also suggest that teacher ratings, observations, informal tests, and their readiness inventory be used as supplementary aids, because ". . . paper and pencil readiness tests do not measure all the components of general readiness for specific skills, such as reading or arithmetic."

The length of each of the subtests is quite short, probably due to the age of the child for which the test is constructed. The writer concurs with the authors' recommendations that the subtest scores not be used and

would further advise that these scores should not be recorded on a student's record. The studies provided in the manual by the company indicate that as predictors the subtest scores are not consistently stable enough from sample to sample to place confidence in them. The total score, however, appears to be a good predictor of reading success.

Reliability

Reliability was determined with odd-even coefficients corrected by the Spearman-Brown formula. Although the sample for which the reliability was computed consisted of students from three school systems, the sample is not further described. The reliability of the total score was above .90 in all three of the sample groups. Reliabilities of the subtests are lower but sufficiently high to merit confidence--except for scale two, the listening section. The reliability of this scale is much lower than one would desire from the data presented in the manual. Since two forms of this test are available, additional reliability information pertaining to the correlations between the two forms should also be reported in the manual. In the section on "construction" of the test, the manual only states that item-discrimination indices were used as a basis for selection of the items for the two final forms that are considered "equivalent." More evidence to support the assumed comparability of Forms A and B is needed. A separate handout sheet does report these data and is intended to be included in the manual in the future.

At the present time, the test manufacturers do not provide in their manual a complete enough description of how the test acts as a predictor for multiple samples. In a new manual, evidence is offered that the test may not serve so well with rural southern children where the correlation between the Metropolitan Achievement Test and the Stanford Achievement Test is .60-.63.

The test manufacturers do provide many single-sheet summaries of data collected by individuals in various locations who have used the test. These data strongly support earlier conclusions about the test. The correlations of the *Word-Meaning* and *Listening* subtests with actual reading success are much too low in many populations for one to place much confidence in the scores as predictors of which children will learn to read. This fact is particularly true of samples from the South (South Carolina and Mississippi) and samples from predominately rural states (Wisconsin). Even the total score is correlated with reading achievement in these samples below what would be desired. The question of the test manufacturer's responsibility in providing adequate data about the validity and reliability of the tests has been raised in this review. It is maintained that before a test is offered for sale adequate information should and must be gathered by those who desire to sell the test. The data which were collected in schools after the test was presented for sale, were interesting but were gathered under such varying test conditions and computed by such a variety of personnel that it is difficult to know how to interpret the matter.

Adequate validity, reliability, and norming procedures should be obtained before the test is presented for sale.

Summary

The test appears to be a good predictor of readiness for kindergarten and beginning first grade youngsters. The test is probably most suited for middle class suburban children. In addition, the total test score may serve with this population as a rough screening device of intellectual functioning. With children from middle class communities outside of suburbia, the test probably will do equally as well. The test scores should be interpreted with great care with lower socioeconomic, rural, and southern areas. The writer strongly advises against the individual use of the subtest scores, as do the test manufacturers, particularly for *Word-Meaning* and *Listening*.

• The Murphy-Durrell Reading Readiness Analysis

Overview

The Murphy-Durrell Reading Readiness Analysis published in 1965 is an outgrowth of the Murphy-Durrell Reading Readiness Test published in 1949. Only one form of the test is available. Subtests include *Phonemes*, *Letter Names*, and *Learning Rate*.

Students record their answers directly on the test booklet, and a separate scoring key is helpful in scoring the test. The directions to the examiner are clear and concise, and the test format is attractive and appears to be easy for children to follow. The pictures and printing are legible and current, and there appears to be a minimal amount of cultural biasing because students do not have to name the items pictured.

One will find tables in the manual for converting raw scores into stanines, percentiles, and quartiles. These tables include conversions for both the subtests and the total test score. The tables are easy to use, and the authors have provided a clear and concise description of how to interpret each type of score and how to plan instruction on the basis of these scores.

Norms

Use of the conversion tables is, however, somewhat limited because of the lack of a complete description of the standardization population. The manual states that several pertinent data items, such as type of community, median income, and number of years of education completed by adults in the community, were collected and utilized in selecting the norm group; however, none of this information is included in the manual The regional distribution of the norming population is also quite uneven. Approximately six percent of the norming population is from the South, and that six percent comes from only one state. Under these conditions, one would probably find the test results more interpretable if one developed one's own local norms rather than relying on the publisher's norms.

Reliability

The reliability of the test was determined by randomly selecting 200 students from the norming population and computing split-half correlations. Basically, this procedure is to divide the test in half by putting all the odd-numbered items in one group and the even-numbered items in another group; the reliability is then the agreement between these two halves. Using this procedure, the total test score appears to be quite reliable. The manual also cautions the test consumer to think of a pupil's score as falling within a range of possible scores rather than at a particular point. This is useful advice for interpreting scores. The reliability evidence would be more useful if a complete description of the 200 cases were given because reliability coefficients can vary from population to population.

Validity

The manual suggests that prospective test users examine the test to determine content validity, but the authors do not include a discussion of their definition of "reading readiness." According to the American Psychological Association's *Standards for Educational and Psychological Tests and Manual*, "... the manual should indicate clearly what universe (content) is represented and how adequate is the sampling." This standard is considered "essential."

Predictive validity evidence for a relatively small sample of 200 pupils from four school systems in Kansas indicates that the Murphy-Durrell test given at the end of kindergarten is somewhat predictive of reading achievement as measured by the Stanford Achievement Test—Primary I when given at the end of first grade. Approximately 43 percent of the performance on the reading test was accounted for by the readiness test. The publisher has also indicated that additional predictive validity data which have been gathered since the publication of the test are available; but if this information is to improve the interpretation of the test scores, it should have been gathered prior to the publication of the test and should be included in the test manual. The lack of description of the population or the reading program for these 200 students in Kansas also limits the interpretation of the validity data.

Evaluation of Subtests and Items

The three subtests are somewhat different from the usual subtests on a reading readiness test. For the *Phonemes* subtest, the student is to select from four words those that begin with a phoneme given by the examiner. The words are represented by pictures, but the item in each picture is named by the examiner. For the *Letter Names* subtest, the child is to select from five alternatives the letter named by the examiner. Part one tests knowledge of capital letters, and part two tests knowledge of lowercase letters. The *Learning Rate* subtest is a measure of the number of words retained an hour after instruction.

All of the subtests are closely related to the actual task of learning to read, and, therefore, they would seem to be quite useful for diagnosing students' readiness to begin reading. The test authors, however, fail to include any discussion of other factors which should be considered by the teacher. The discussion of the use of the test results seems to indicate that the skills measured by the test are the only factors to be considered.

For a random sample of 200 cases from the norming population, the split-half reliabilities of the *Phonemes* and *Letter Names* test were quite high (.94 and .97), but the *Learning Rate* test had a reliability of only .88; this result is probably partly caused by the relatively short length of the test. It is believed that these reliability indexes are high enough for one to make separate use of the subtests.

The use of the subtests for predicting reading achievement is lower than the total test score. *Phonemes* is the best predictor, and *Learning Rate* is the poorest. The range of these predictions indicates that from 15 to 58 percent of beginning reading achievement is accounted for by the various subtests when they are considered separately. The authors suggest that the total score should be utilized in planning instruction. Because of the lack of diagnostic validity evidence the writer would support this procedure and suggest only limited use of separate subtests.

Summary

The Murphy-Durrell Reading Readiness Analysis includes subtests which are very similar to the skills of beginning reading. However, the skills are related primarily to the decoding aspects of beginning reading, and it is, therefore, suggested that if this test is going to be used as an indication of readiness, other factors should be considered. The test is probably most useful as a criteria measure of specific skills, and use of the test should probably be based on one's classroom experience and the development of local norms.

Chapter 3

SELECTING A READING ACHIEVEMENT TEST

PROBABLY more tests are administered at the end of the school year than the beginning, and this practice is a reflection of the purposes a teacher has in mind in using tests: that is, most schools are more concerned with how much a child has learned in a given year than in giving a test early to assist in planning instruction.

Tests should be selected according to the desired purpose. If one wishes a quick overview of one's students' current status, administering an achievement test early in the school year is a desirable practice. If one wishes to evaluate the instructional program, administering the achievement test late in the school year is a useful practice. To determine the specific strengths and weaknesses students possess, a diagnostic test or an achievement test with appropriate subtests should be selected.

These reasons should guide one's selection of a reading test. One will want to know how accurately the subtests can be used as diagnostic assessments, how reliable and valid the results are, and exactly how the test was normed in order to determine whether the grade placement, percentiles, or stanine scores are appropriate for one's class. In essence, these points should be considered in choosing the test. The charts in the Appendix give a quick overview of the major characteristics of the tests reviewed and an evaluation of each test.

• California Reading Tests

Overview

The California Reading Tests are part of a larger battery of tests called The California Achievement Tests (CAT). These tests have a long history and have been through several revisions since they first appeared in 1934. This review covers the current 1957 edition renormed in 1963.

The reading tests are divided into three levels: lower primary, grades 1 & 2; upper primary, grades H-2, 3 & 4; and elementary, grades 4 through 6. In addition, several forms of the test are provided for each level. It should be noted, however, that this review covers only Form W as the test manual and technical report do not include data for any other form. The test manufacturers rightly caution against using other forms for research purposes and clearly state that the standardization and normative work were done only for Form W. The publisher states that critical users or

experimenters should use Form W ". . . since conclusive evidence of form equivalence is not available" (page 24, CAT, *Technical Manual*). While there is no empirical evidence that the forms are not equivalent, there is more importantly no evidence that the forms are equivalent. Thus, all forms other than Form W are an unknown quantity and are not recommended for use.

All of the reading tests are divided into two subtests, *Reading Comprehension* and *Vocabulary*. Each of these subtests is then divided into sections. The manual suggests that the sections were devised to ease administration and to reduce the time that it would take the child to complete the entire test. The section scores are *not* to be used for grouping or instructional purposes.

The California Reading Test, Lower Primary level, is divided into two subtests: *Reading Vocabulary* and *Reading Comprehension*. The *Reading Vocabulary* section contains four sections with a total of 75 items: Word Form, 25 items; Word Recognition, 20 items; Meaning of Opposites, 15 items; and Picture Association, 15 items. The 75 items seem to be an adequate measure of reading vocabulary. Item 12, Test 1, Section D, is the only item that needs to be questioned in terms of picture-response clarity. It is the reviewer's opinion that the illustration could be misinterpreted by examinees. However, this is a minor point.

The *Reading Comprehension* section is divided into two parts: Following Directions, five items; and Interpretation, ten items. The items in these two parts are reportedly designed to measure skill in following directions, noting specific facts, and making inferences. Although these three skills appear to be what the authors are attempting to measure, it was difficult for a group of experienced reading teachers to ascertain which reading skill was being tested in any particular item in the *Reading Comprehension* subtests. The test authors recommend that the section scores be used as indicators of areas of reading disabilities. However, the length and reliability of the various sections are such that attaching much significance to the vocabulary or comprehension section scores individually should be avoided. The various sections appeared to be controlled for readability. The items are arranged from easy to difficult in each section, and content of the test is current and does not appear to favor specialized backgrounds. A separate section, *Letter Recognition*, is included at the end of the regular test and is to be used with those students who obtain very low test scores on the *Reading Vocabulary* and *Reading Comprehension* Tests. The *Letter Recognition* section contains 24 items and is designed to help the teacher gather additional information on general performance with verbal symbols. This section requires that the examinee identify alphabetical letters in their capital and lower-case forms. The child indicates whether the words joined by a dotted line are the same or different. The two words may or may not appear in the same printed form. This particular section could yield valuable information on certain word recognition skills for poor readers.

For the most part, the names of the *Reading Vocabulary* sections adequately describe what each section is attempting to measure. In the writer's opinion *Reading Comprehension* does not adequately describe the nature of the tasks in this section of the test.

The California Reading Test, Upper Primary level, is divided into two tests: *Reading Vocabulary* and *Reading Comprehension.* The names of the *Reading Vocabulary* and *Reading Comprehension* tests are adequate for a general description of the task involved. *Reading Vocabulary* contains two sections with a total of 45 items: Word Recognition, 20 items, and Meaning of Opposites, 25 items. *Reading Comprehension* contains three sections with a total of 51 items: Following Directions, 15 items; Reference Skills, 15 items; and Interpretation of Material, 21 items. A separate *Word Form* test which consists of 25 items is included at the end of the regular test for diagnostic purposes and is to be used only with those students obtaining very low scores on the total reading test. The *Word Form* test consists of pairs of words which the examinee is to mark as same or different in appearance. In a very limited way, the *Word Form* test provides the examiner with some diagnostic information in the word recognition skills area. The test authors claim that the principal value of the section scores is their indication of existing weaknesses. However, the sections are relatively short, and the attachment of much significance to individual vocabulary or comprehension scores would not be warranted by the reliability coefficients. Readability appears to have been a consideration in the construction of the various subtests. Item progression in each subtest is from easy to difficult. The content of the test items is current and does not appear to favor any particular background of experiences.

The California Reading Test, elementary level, is divided into two parts: *Reading Vocabulary* and *Reading Comprehension.* Mathematics, science, social science, and general vocabulary sections are included under *Reading Vocabulary.* Following directions, reference skills, and interpretations sections are included under *Reading Comprehension.* The tests names do adequately describe the task in each of the sections. Fifty vocabulary items are used for the four vocabulary sections, and 60 are used for the comprehension subtests. The reviewer noted that experienced teachers, when asked to do so, had difficulty classifying the 50 vocabulary terms into the categories of mathematics, science, social studies, or general vocabulary. A similar group of trained reading teachers disagreed in labeling the skills the authors claimed were being tested in various items in the *Reading Comprehension* section. The test authors caution against attaching undue significance to the scores on separate sections but go on to state that the principal value of the scores is their indication of existing weaknesses. There is no evidence cited to support this claim, and the writer recommends using the scores with caution. The sections do appear to be controlled for readability, and the items are arranged from easy to difficult. The item content is current, and favoritism for specialized backgrounds has been avoided in item development.

The administration of the California Reading Tests, all levels, appears not to be difficult. The directions and examples are clear and concise. All the subsections are timed tests; therefore, a stopwatch or clock with a second hand is needed. Scoring time is less rigid for the Elementary level test. Students may mark their answers on separate answer sheets, such as IBM, or use SCOREZE, a self-scoring answer sheet available through the firm. IBM answer sheets may be less expensive and can be scored rapidly by machine or using hand scoring overlays.

For the Lower and Upper Primary level tests, students are required to mark their answers in the booklets, and these must be scored by hand. When scoring any test by hand one is cautioned not to make a diagnostic interpretation of these subsections. These scores are not to be used in this way, and faulty interpretation of a student's skills can result from interpreting individual items.

Grade placement, percentile stanines, and standard scores are provided for each test and total score. All tables are clearly identified and easy to use.

Although the manual cautions that the test sections are not normed and no grade placement should be attached to these scores, there are columns on the profile sheet for obtaining these scores. The technical report rightfully cautions their use. One suggests that the section scores should not be converted to grade equivalent scores and that the test manufacturers remove those columns from the profile.

Although a Diagnostic Profile is provided, one does not suggest its use. Data are offered in the technical report and the manual for the tests as general reading achievement measures, and the power of these tests to diagnose specific reading difficulties has not been demonstrated. Until such evidence is available the writer suggests only the use of the three scores—reading vocabulary, comprehension, and total score as a measure of general reading achievement.

Norms

The norming population for the California Achievement Test Primary, Upper Primary and Elementary, 1957 edition, was extensively controlled for geographical and instructional program bias. The renorming of the test in 1963 took a more limited sample but appears to be quite adequate in number of pupils and range of abilities. The technical report discusses the norming extensively and presents cautions in interpreting individual scores. As in other reviews, the establishment of local norms is suggested.

Reliability

Split-half reliabilities are reported only for the two subtests and total scores for each level. The reliabilities reported are for one grade level for each test; that is, 1.7 for Lower Primary, 2.7 for Upper Primary, and 5.1 for the Elementary.

The Lower Primary level reliability for *Vocabulary* is sufficiently high, but the *Reading Comprehension* is .77, somewhat lower than desired. The total score is .88. The Upper Primary and Elementary level reliabilities are very acceptable.

Reliabilities for each grade level rather than those reported would be desirable. Due to the fact that the Lower Primary is for first grade and beginning second, a sample of 1.7 grade level is adequate. However, the Upper Elementary is designed for fourth, fifth, and sixth grades, and reliabilities are only reported for grade five. It is assumed fourth and sixth grades would have similar reliabilities, but one would like these data reported.

The Upper Primary test-retest reliabilities reported for 90 students at grade placements of 2.8 and 3.8 are lower than one would desire for the reading comprehension section (.59). The vocabulary and total score test-retest reliabilities are sufficiently high.

Test-retest reliabilities are reported for 90 students in grades 4-8 and 5-8 and 125 students in 5-8 and 6-8. These are quite high and well within a desirable range.

In addition, data are reported for all three levels on test-retest reliabilities of students who took the lower form and were then tested on the next highest form. These are all adequate except for the correlation of the Reading Comprehension section of the Upper Primary to the Elementary Reading test of comprehension where the coefficient as reported is .54. Why the Upper Primary Reading Comprehension test-retest reliability for the same form or next highest form is so low is not made clear.

The lower reliability may be a product of rapid skill development of children at this age or it may reflect program changes in the teaching of reading. Comprehension skills tend to be stressed by teachers at the end of second grade and more in third grade. This score may be a reflection of changes in program emphasis, but this is only a guess. One should interpret this score with other information available from the classroom program, such as the pupils' daily reading assignments and independent reading activities.

Validity

Validity for the series is reported in two ways: 1) by having a team of experts inspect all items and 2) by correlating the CAT with other achievement tests.

The "experts" who examined the items generally agreed that the items included in the test were ones that were essential concepts or of major importance. As previously suggested, any test must be examined by the teacher to see if it matches the content as taught in his classroom. Examination of the test leads one to agree that items on the test are important aspects of a basic reading program.

Listed in the manual, a table called "Diagnostic Analyses of Learning Difficulties" breaks the items into groupings within the test. One should

examine this table to determine agreement with what is included. The writer can find no empirical evidence to suggest these skills are being measured; however, no criticism of the test authors is made for this lack as an achievement test must sample from a range of skills. One does caution, however, that evidence to support the test as a diagnostic instrument is not given and further, that the manual and technical report caution against using it so. There is no evidence offered, for example, that the mathematics vocabulary is a reliable or valid sample of mathematics vocabulary. This statement can be applied to any other of the subsection scores. Their inclusion on the profile may mislead one into confidence that the child has mastered these skills. These scores are, as the test authors suggest, only "cues" for one to verify with other data.

The correlations of the CAT with other reading achievement tests are high but none are reported for the Lower Primary.

Extensive correlational data among the subtest scores and the California Short Form Test of Mental Maturity, 1963 revision, are reported. These data can be interpreted as validity of the test. Again, Lower Primary Reading Comprehension scores correlated much lower than the other subtests.

The writer questions, however, that the correlation between the achievement test and the Mental Maturity Test should be this high. The correlations as reported suggest that either the Mental Maturity Test is an achievement test or that the California Reading Test is a mental maturity test. The correlation of the Upper Primary and Elementary Reading total score and the CTMM-short form are .79 and .81, respectively, indicating that these two tests measure somewhat the same skills.

Summary

The California Reading Tests span the elementary level nicely, are easy to administer and use, and are in general reliable and valid measures of reading behavior. One will have to inspect the content of the test to determine how closely it matches one's classroom instruction. The Lower Primary Comprehension test is not so reliable as one would desire, but the three scores, (vocabulary, comprehension, and total) for each test are of use as a measure of general reading comprehension.

The major criticism of the test is that the section scores on the profile sheets and the Reading Diagnostic Profile have not been demonstrated to be reliable or valid measures. Although the manufacturer cautions against the use of these scores, their inclusion, it is believed, can be misleading.

Only information on Form W is provided, and the writer suggests that this is the form to be used until information is provided by the publisher for the other forms.

• Gates-MacGinitie Reading Tests

Overview

The Gates-MacGinitie Reading Tests are a new edition standardized in 1965 and developed to replace the Gates Primary, Gates Advanced

Primary, and the Gates Reading Survey. Included in this seven-test series are tests for all grade levels, from first to twelfth. This review will cover only the five tests from grade one up to and including grade six.

These five tests are Primary A for grade one, Primary B for grade two, Primary C for grade three, Primary CS for grades two and three, and Survey D for grades four through six. Primary A, B, and C and Survey D include subtests for *Vocabulary* and *Comprehension;* in addition Survey D has a *Speed and Accuracy* subtest. Primary CS is a *Speed and Accuracy* test for grades two and three. Two forms of the test are available for Primary A, B, and C: three forms are available for Primary CS and Survey D.

The scoring of all the tests is aided by scoring overlays. Tables in each manual are provided to convert subtest and total test scores to grade norms, standard scores, and percentiles. There are tables for beginning, middle, and end-of-the-year testing times for each grade except first. Middle and end-of-the-year norms are available for first grade. These tables were developed by norming the test at the beginning and end of each grade and then interpolating to estimate middle-of-the-year norms. This procedure is far superior to the usual practice of administering a test at only one time during the year.

In the technical manual the authors also provide tables for interpreting differences between subtest scores and also for evaluating differences between scores in estimating reading growth. The comparisons of subtest scores are based on standard scores.

The educational significance of the differences of these scores is determined by the probability that two scores would differ by a certain amount fifteen times out of a hundred. If these subtest differences are apt to occur more than fifteen times out of a hundred, this information is considered to have educational value in planning a reading program. Formulas are also provided for determining the significance of average subtest score differences for groups of children.

The tables for evaluating the significance of reading test gains also utilize standard score differences, and these differences are again considered significant only when they are apt to occur more often than fifteen times out of a hundred. The use of these tables will be very beneficial in interpreting the test scores. The technical development of the tables appears to be very satisfactory, and the suggestions for their use by the test authors are excellent.

For all of the tests, the total reading score is determined by averaging the standard scores of the subtests. This average standard score can then be converted to a percentile or grade score. The test authors correctly point out that, when determining averages, it is not good practice to add and divide raw scores because they are not based on an equal-interval scale.

The Gates-MacGinitie Reading Test, Primary A for grade one and Primary B for grade two are quite similar. Each has two subtests, *Vocabulary* and *Comprehension.* The *Vocabulary* subtest has 48 items. The student is

to match a picture with the word it represents; four alternative responses are provided. Some of the items appear to be measuring visual discrimination of words, and others seem to measure the student's ability to determine the *meaning* of the picture. The authors suggest that the test measures ability to recognize isolated words. The pictures are clear and up-to-date, and there seems to be a minimal amount of cultural biasing in the selection of items.

The *Comprehension* subtest measures the student's ability to read and understand whole sentences and paragraphs. The student is to match the sentence or paragraph to one of four pictures. For some of the items, it appears as though the student could determine the correct response from reading only one or two words in the selection. Because of this condition, it is probable that the test is not measuring a much different ability than the *Vocabulary* subtest. The reported correlations of these subtests (.67 for Primary A and .78 for Primary B) would seem to support this contention.

Primary C for grade three follows the same pattern as Primary A and B. However, the *Vocabulary* subtest has a total of 52 items; for 12 of these, the student is to match the correct word with a picture, and for the remainder of the items he is to select the best synonym for a stimulus word. The *Comprehension* subtest includes 24 paragraphs each of which is followed by two multiple choice questions. Some of these questions ask students for meanings of words in the paragraphs and, therefore, as might be expected, the correlation of the subtests as reported in the technical manual is .83.

Primary CS for grades two and three is a test of reading speed and accuracy. There are a total of 32 items on the test. Each item includes a short paragraph and a five-option multiple choice question or incomplete statement. The students are given a total of seven minutes to work on the test. Two scores, number attempted (speed) and number correct (accuracy), are determined for each student. According to the publisher, the accuracy score correlates .78 with both the *Vocabulary* and *Comprehension* subtests for Primary C. The speed score correlates .54 and .53 with *Comprehension* and *Vocabulary* respectively. From these correlations it appears that the accuracy score is measuring the same set of skills as the *Vocabulary* and *Comprehension* tests are but the speed score appears to be measuring a different variable.

Survey D for grades four, five, and six includes *Vocabulary* and *Comprehension* subtests and also the *Speed and Accuracy* subtest, similar to Primary CS. The *Vocabulary* subtest has 50 items which measure the student's ability to choose the best synonym for a stimulus word. Each word is presented in isolation, and the student is to choose the correct response from five alternatives. The test is timed; however, it seems likely that most students should be able to complete the test in the 15 minutes allowed. The *Comprehension* subtest consists of 21 paragraphs in which there are a total of 52 blanks. For each of these blanks the student is given five alternatives to choose from in selecting the word that best fills the blank. Students are allowed only 25 minutes to work on this subtest. The

Speed and Accuracy subtest follows the same format as Primary CS. However, there are 36 items, and the students are given five minutes. The use of a time limit for the *Speed and Accuracy* subtest is, of course, necessary; but it does not seem to be a defensible procedure on the *Comprehension* and *Vocabulary* subtests where the attempt is to measure reading power. While it seems probable that scores on the *Vocabulary* and *Comprehension* subtests would not vary significantly if more time were allowed, evidence of this condition should be provided by the test publisher.

The directions for all of the tests and the various subtests are clear and concise. The subtests follow a logical pattern of increasing difficulty of items. Some test authorities have suggested that it is better practice to intersperse difficult items with easier items, but there is contradictory evidence as to which is the better practice. The use of differences in subtest scores should be interpreted cautiously; the tables provided for this purpose in the technical manual should be utilized. The high correlations between the *Vocabulary* and *Comprehension* subtests indicate that these subtests are measuring quite similar traits; it also appears that the *Speed and Accuracy* subtest is measuring a somewhat different trait.

Norms

The tests were normed on 40,000 pupils in 38 communities. Because of the number of test levels and test forms it is probable that each test was normed on about 2,500 pupils. The authors state that the communities for the norming population were selected on the basis of size, geographical location, average educational level, and average annual income. Despite the fact that these variables were allegedly controlled, the authors do not describe the population. The norms can be cautiously accepted as being representative of national performance; however, for a more precise and meaningful interpretation it would be best to develop local norms.

Reliability

Reliability indexes were computed by both the split-half procedure and the test-retest procedure utilizing different forms of the test. Reliabilities are reported for each subtest at every grade level. These reliabilities are based on testing in five separate communities, but these communities are not further described. A more complete description of these communities is vital in interpreting the reliabilities for one's classes. In general, the reliabilities are high enough for one to feel fairly certain that the score a student will receive on one form of the test on one day is likely to be the same as the score he receives on another form on another day. The alternate form reliabilities for the *Vocabulary* and *Comprehension* subtests range from .81 to .89, but the *Speed and Accuracy* subtest only ranges from .67 to .86. The split-half reliabilities for *Vocabulary* and *Comprehension* range from .88 to .96. The split-half reliabilities for *Speed and Accuracy* are not reported because of the problems of correlating alternate halves of a stringently timed test.

Validity

Validity evidence for the test is very limited. The test appears to have face validity for measuring what it purports to measure; and, according to the test authors, the items were selected on the basis of a tryout with more than 25,000 pupils. There is, however, no description of the curriculum content which this test is supposed to be measuring. There is also no evidence that the subtests were selected by examining the content of reading programs. However, the authors' accepted expertise as specialists in the area of reading behavior somewhat diminishes this criticism. As with other tests, it is suggested that if one decides to use any of these tests with specific students, one should carefully examine the objectives of one's reading program and compare these to the content of the test.

Correlations between Survey D subtests and Lorge-Thorndike Verbal IQ scores are reported for grades 4, 5, and 6. These correlations indicate that for all of the subtests the similarity between the vocabulary and the verbal IQ scores become closer at higher grade levels. In addition, it appears that vocabulary and comprehension scores are more related to verbal IQ than are speed and accuracy scores. These correlations lend support to the general conclusion that there is a great deal of similarity between group measures of verbal IQ and group measures of reading achievement. Most of this similarity is probably due to the amount of reading that is necessary on a group verbal IQ test.

Summary

The Gates-MacGinitie Reading Tests provide a measure of general reading achievement for students from grades one through twelve. Only those tests used in grades one to six are included in this review. In general, the tests are well constructed, and the authors have provided a useful procedure for interpreting the differences between subtest scores. This is a welcome trend in the development of reading tests.

The tests have been normed at both the beginning and end of each grade, and the subtests and total test scores are quite reliable. One should certainly examine the test's validity for measuring the objectives of a specific program by comparing the program objectives to the test objectives. The development of local norms also would aid in the interpretation of test scores. This test series appears to be one of the better instruments available for measuring the reading achievement of students. It should be useful for evaluating growth, screening students who are in need of more diagnostic testing, organizing instructional groups, and cautiously diagnosing subskill deficiences.

• Iowa Silent Reading Test

Overview

The Iowa Silent Reading Tests (new edition) are available in two separate tests, one for grades one to eight and the other for grades nine to thirteen. This review will consider only the elementary level of the test. Many of the shortcomings of the test are due to its age. The test booklet

was published in 1942, and the latest test manual copyright date is 1943. There are four forms of the test available. Eight separate subtests are included: *Rate, Comprehension, Directed Reading, Word Meaning, Paragraph Comprehension, Sentence Meaning, Location of Information Alphabetizing,* and *Location of Information—Use of Index.*

Hand scoring of the test is somewhat difficult because the test booklet must be turned upside down in order to score some of the subtests—an unnecessary complication for pupils taking the test. Because of the hand scoring difficulty, it is suggested that if one uses this test, one should utilize the machine-scored answer sheets which are available from the publisher.

Raw scores must be converted to standard scores before being converted to percentiles, grade equivalents, and age equivalents. The total standard score is determined by computing the median standard score for all the subtests. A profile is printed on the front of the test booklet for comparing subtest standard scores. It is recommended that one does not use this profile for diagnosing of students' reading abilities for two reasons: 1) the norming population is quite inadequately described and one would be comparing one's students to some unknown group; and 2) the reliabilities of several of the subtests are quite low. For example, the reported split-half reliability of the comprehension subtest is .68 for grade six students.

The *Rate* subtest has two serious weaknesses. First, the students are not given any purpose for reading the material; they are told only to read carefully so they can answer questions about the story. The writer's belief concerning reading-rate tests is that the test should measure how rapidly a student can accomplish a specific purpose. Secondly, the directions state that the student may not look back at the selection to answer the questions, a stipulation meaning that the test is very heavily loaded with an immediate memory ability.

The *Directed Reading* subtest appears to be an attempt to measure the student's skimming ability. However, the use of format and typographical aids in the selection would greatly increase the value of the test. The *Word Meaning* and *Paragraph Comprehension* subtests follow the traditional pattern of utilizing words in isolation and multiple choice questions following a selection. Use of these subtests should be based on an analysis of a reading program and the content of these subtests.

The *Sentence Meaning* subtest appears to be measuring knowledge other than reading ability. For example, one statement asks, "Do most children attend the public school in the summer time?"

The *Alphabetizing* and *Use of Index* subtests are designed to measure reading-study skills. It does seem that if the measure of reading-study skills was desired, the authors should have included measures of other skills such as use of the Library and using parts of a book.

Norms

Norms for the test were gathered in the spring of 1942 and are based on 9,000 pupils in "... 19 communities in 13 states widely distributed

geographically." Due to the very limited description of the norming population and also because the test was normed over 25 years ago, there is absolutely no validity for these norms and one should not use them under any circumstances. The test could be used for comparing student growth if one were to develop local norms.

Reliability

As indicated previously, the reported split-half reliabilities for most of the subtests are too low to be sure that a student's score will not vary considerably from day to day. The total reading score, which is the median standard score for all the subtests, is more reliable. However, because the reliabilities are based on a poorly defined norming population, an evaluation of the reliability of the test is very difficult.

Validity

The test has been defined by the authors as a reading-study skills test, and the make-up of the subtests appears to have face validity. The development of the outline of skills for the test was based on a textbook, *Measurement and Evaluation in the Elementary School*, which one of the authors of the test coauthored. The only other validity evidence for the test is the report of a small study which indicates that most of the subtests are only minimally related to one another. This factor would aid in the diagnostic use of the subtests if 1) the subtests were more reliable and 2) the population for the tryout consisted of a better described sample at more than one grade level.

Summary

The Iowa Silent Reading Test—Elementary Edition will be of only very limited value for use with one's classes. The skills measured by the test appear to cover a broader range of skills than most elementary reading tests. But, the inadequacies of the norms, the antiquity of some of the items, the lack of validity evidence, and the limited reliability of some of the subtests should cause one to reject it.

• Metropolitan Achievement Tests—Reading

Overview

The reading tests of the Metropolitan Achievement Series are part of a larger battery of tests which span the last half of first through the ninth grades. This review will cover the reading tests of the first through sixth grades.

The series dates back to 1932 and has been revised several times, the most recent being the 1959 edition which is the one used in this review. There are four levels of tests covered in this report: Primary I to be used in the latter half of grade one or beginning grade two, Primary II for use in grade two, Elementary for use in grades three and four, and Intermediate for use in grades five and six. An advanced form (for grades 7-9) is available also but is not reviewed here.

There are three forms available for Primary I and II and four forms for the other levels. However, neither the "Manual for Interpreting" nor the "Directions for Administering" present data to indicate form comparability. A separate summary is available from the publishers reporting correlations between forms A and B from a single school district. These intercorrelations are high for all levels of the reading tests. The manual does not make it clear whether the other data presented—for example, the split-half reliabilities—are for all forms of the test or for just one. It is, therefore, suggested that forms A and B appear to be comparable. They may be used with confidence until such time as other data are reported.

Grade placement, percentiles, and stanines are provided for all levels, and instructions on how to compute local stanines are given in the Manual for Interpreting. The Directions for Administering are short and easy to read, but they do not give sufficient information for evaluating the tests. Both documents are needed to evaluate the tests.

For each level there is more than one test. Primary I and II and Elementary contain *Word Knowledge, Word Discrimination* and *Reading Tests* subtests, while Intermediate contains *Word Knowledge* and *Reading Test* subtests. The Primary I battery is divided into a 35-item *Word Knowledge* test which takes 15 minutes, a 35-item *Word Discrimination* test which takes 12 minutes, and a 45-item *Reading Test* which takes 35 minutes.

The directions are clearly written and easy to follow. A watch or clock with a second hand should be available for the timed sections. The test is designed to measure orally presented words, student sight vocabulary, and the student's ability to comprehend sentences and paragraphs. The titles seem descriptive of what is required. *Reading Comprehension* refers to a limited number of skills, and one should examine the test as the test manufacturer suggests, to determine whether the skills taught in one's program are being measured.

One sample item is given for each test. Pupils who have limited test experience may have difficulty following directions with only one example. However, the reliabilities seem sufficiently high to indicate this problem may not occur.

The pictures are clear, and one-line drawings are such that they should minimize the student's becoming confused by too complex pictures. The content appears to be current and not to favor a special population. One study provided by the test publisher did indicate, however, quite low test-retest reliabilities (.67) on forms A and B from June to September for Negro boys on the *Reading Test*. These data may be influenced by the summer lapse. However, interpretation of results with this group should be done with care.

The publishers point out that the norming was done with groups of pupils of average and slightly above average mental abilities. This fact should be taken into account in interpreting student scores.

The Primary II Reading tests contain a 37-item *Word Knowledge Test* which takes 18 minutes, a 35-item *Word Discrimination* test which takes 12 minutes, and a 51-item *Reading Test* which takes 35 minutes.

The *Word Knowledge* subtest is divided into two sections: the first asks the child to select the correct word to match a picture; the second requires the child to complete a sentence with four alternatives. The example question is for the first section, and no example is given for the second section. This omission may cause some confusion among young children taking the test. The directions appear to be easy to follow; the line drawings are clear and current, and the items are spaced well on the page.

The *Word Discrimination* subtest attempts to measure auditory and visual discrimination ability. The teacher pronounces one word, and the examinee is directed to find the graphic representation of that word among the four alternatives provided. To make sure the examinee does not misinterpret the stimulus word, the examiner presents the stimulus word in oral context. The stimulus words and alternatives were chosen carefully. Words containing a variety of consonants, vowels, blends, and digraphs are represented. Format, appearance, type size, directions, and time allotment are adequate for the population intended.

The third subtest, *Reading*, is divided into two parts—sentence reading and story reading. There are 13 items in sentence reading, with each item containing a line drawing and three sentences. The examinee demonstrates his understanding by choosing the sentence that best describes the line drawing. The line drawings are the same size as those used in the *Word Knowledge* test. The validity of the sentence reading section of the *Reading* subtest must be questioned until evidence is offered as to what this test is measuring. All tests must select from the range of skills involved in being able to read; the writer does not expect a general reading test to measure everything. He does ask, however, that evidence be offered to support the contentions that the publishers make about their tests. The story reading section of the subtest contains 10 passages; each is followed by a number of multiple choice questions. The manufacturer states that the questions test main ideas, details, inferences, and specific word meanings. No questions were asked on skills relating to organizational ability. In the children's score box for the test, the *Reading* test is divided into scores for Sentences, Stories and Total. No evidence exists to support converting these section scores into grade equivalent, percentile, or stanine scores. Only the total *Reading* scores should be used, and it is suggested that the manufacturer remove these columns from the score box.

The Metropolitan Achievement Test, Elementary Reading Test, contains two subtests: *Word Knowledge* and *Reading*. The first subtest title is descriptive of what specifically is being measured, but the second is too broad to serve much usefulness as a subtest title. *Word Knowledge* consists of 50 items, each of which is composed of a single vocabulary word placed in partial context and followed by five words, one of which correctly completes the context.

The vocabulary words being tested appear to be carefully selected and representative of third and fourth grade children's vocabularies. Each of the four alternatives appears to have been carefully chosen and matched within each item. The subtest progresses from relatively easy words to more difficult words. The directions and format are easy to follow and

should not be confusing to either examiner or examinee. The time limit of 15 minutes for the 50 items in the subtest appears to be realistic in terms of third and fourth grade children's reading ability and speed.

The *Reading* subtest consists of nine passages, each of which is followed by a number of multiple choice questions. The questions appear to be measuring main ideas, details, inferences, and specific word meanings. No questions were asked on skills relating to organizational ability. Because the subtest title *Reading* is so broad, it is difficult for the reader to determine what should be included in this subtest in order to adequately appraise the validity of the subtest. The time limit of 22 minutes for reading the nine passages and answering the 44 related questions appears to be adequate in terms of the reading speed of third and fourth grade children.

Readability of passages and questions seems appropriately controlled. The nine passages are logically arranged from easy to difficult. The questions are clear and concise; however, those questions testing independent word meaning could be improved. The examinee is asked to select an appropriate definition for a word used in the context of the story, but the word in the story is not highlighted to facilitate locating the word by the examinee. Since the word is not highlighted, the examinee must use precious time in skimming through the passage to find the word. The directions and format are easy to follow and should not be confusing to either examiner or examinee. The content of the items is not dated, nor does it appear to favor specialized backgrounds of experiences. Both subtests appear to be long enough to provide the examiner with usable results.

The Metropolitan Intermediate Reading Test has two subtests, *Word Knowledge* and *Reading*. Only the title of the first subtest is adequately descriptive. The *Word Knowledge Test* contains 55 items. Each item is designed to test the knowledge of a word judged to occur frequently in children's reading material. Each word selected for the test is presented in a minimal context. The item is completed by the examinee by selecting a single word from five alternatives. Alternatives in each item are carefully matched.

All items seem well chosen with reasonable alternatives. Fourteen minutes are recommended for completion of the 55-item test. This time limit requires the completion of about four items per minute or reading at approximately 56 words and symbols per minute. Neither of these seems unrealistic for fifth or sixth grade children.

The second subtest, *Reading*, consists of seven passages each of which has been carefully graded by controlling vocabulary, sentence length, sentence structure, and overall length of passages. Each passage is followed by a series of multiple choice questions. The analysis of the questions made by the reviewer indicated that the questions attempted to get at main ideas, details, inferences, and individual word meanings almost exclusively. The most noticeable of the missing questions were those attempting to get at organizational ability. Twenty-five minutes' time is recommended for the completion of the seven passages and 44 related questions. Approximately 2,000 words and symbols must be read during the 25 minutes, a

task which requires a reading speed of approximately 80 words per minute. This speed should not be too demanding of fifth or sixth grade students. Both subtests are arranged in a simple-to-difficult order, and both appear long enough to provide reasonable results. The length of the two subtests in the Metropolitan Intermediate Reading Test is at least commensurate with similar subtests found in other reading tests.

The directions for administering and scoring the Metropolitan Intermediate Reading Test are concise and complete. The clarity and language level are appropriate for the grade levels intended. The use of color for underlining key vocabulary words in the *Word Knowledge* subtest, item numbers, and distractor numbers further facilitate understanding and ease of administration. The format, print size, legibility, and currentness of the item content are adequate for the grade levels intended. Readability seems to have been carefully controlled throughout the test. Some of the answers to the multiple choice questions appearing in the test booklet are too crowded for efficient use of the hand scoring key provided. The hand scoring key consists of pieces of cardboard with holes punched where the correct answer is to appear when the answer key is overlaid upon the test booklet. More than one answer could be seen through the holes with the copy the writer possessed. This condition is likely to add to the confusion of using such a scoring device and may even result in an occasional error if the examiner is not exceedingly cautious in his scoring. The hand-scoring procedure is not recommended by the reviewer. An additional set of directions is provided in the manual of directions for use with separate answer sheets.

Norms

The norming population for the Upper Primary, Elementary, and Intermediate Reading Tests consisted of a random sample of 25 percent of the 500,000 students from 225 school systems in 49 states administering the entire Metropolitan Achievement Test in October 1958. The sample was controlled for age to insure normal grade placement of those in the sample. As mentioned previously, the authors indicate that the norms are slightly higher than would be expected with an unselected group. The procedures used by the authors adhere to the rules and constraints of norming. The norms do not include contributions by repeat students and thus will present norms that may be unrealistic for some schools. Sex and socioeconomic data are not available, but further information regarding the geographical distribution of participating schools is available upon request from the publisher. Such data should be obtained by teachers prior to purchasing tests so that test data will be of maximum use.

Validity was established by identifying the reading skills and the levels at which they were tested from reviewing the related research and from examining reading programs. The tests were then constructed to measure the reading skills at various reading levels. No bibliography of the research reviewed is provided, and the combined term "reading programs" is undefined. The description of validity is discussed under "curricular considerations" and is only one paragraph in length. Besides being inadequate

in length and content, the description is vague and raises more questions than it answers.

Additional information on validity is found in the "Manual for Interpreting." The discussion is a complete description of the general nature of the problems of establishing validity but is an incomplete description of the actual validity of the reading tests in the series. An analysis of eleven basal reading series was used from the New York City Board of Education vocabulary study. A careful control was made in placing words at the median level as found in the basal readers. The authors state that "extensive experimentation showed" that the sentences as chosen would not invalidate the test results. What the nature of "extensive experimentation" was is difficult to determine.

Major criticism of the "Manual for Interpreting" and the "Directions for Administering" is that they are incomplete in what they offer. Most information about the test characteristics are included in summary paragraphs for all grade levels in the battery. The test authors make statements about what they believe the test to be but do not make explicit the source of their data other than in general terms. They do rightfully caution not to try to combine scores or to use individual items to interpret a pupil's progress. However, it would take a teacher several hours of reading a great deal of material to find the information he needs to critically evaluate the test and to note the necessary cautions in interpreting the test scores. It should be noted, however, that the material provided by the publishers contains a great deal of analysis and work, consequently making the test a very useful one to measure general reading achievement.

Reliability

Reliability was determined by the split-half technique. Four independent estimates of reliability were made for each test, and the ranges and medians of the four are reported below. Each estimate was chosen to typify a different performance level on the test. One hundred subjects at grade level 3.1 for the Upper Primary Reading Test, grade level 4.1 for the Elementary Reading Test, and grade level 6.1 for the Intermediate Reading Test were randomly selected from each of four school systems to participate in the reliability studies. A total of 400 students participated in the reliability studies for each test. All the reliabilities are high and quite satisfactory for all levels of the test, being .90 or above. No correlations of the Metropolitan series and other reading tests are presented in either of the two documents.

It appears that the validity of the reading tests is based on readability analysis using the Lorge and Flesch formula and the reliability data. Other data to support the test author's claims are desired. It should be noted that extensive support is presented for the *Spelling* test. Similar support is desired for the *Reading* subtests.

Standard score stanine, percentile ranks, and grade equivalent tables are available for score interpretation. The manual of directions provides an outstanding discussion of the merits and limitations of each in an attempt to indicate how the test scores may be used to improve the services of the

school to the child. The section "Use of Results" provides a number of purposes for which the obtained data might prove beneficial to classroom teachers, principals, administrators, and supervisors.

Summary

The Metropolitan Test Battery includes a range of levels of reading tests which should be very useful to measure general reading achievement. All the tests appear to be very reliable. The *Word Knowledge* and *Discrimination Vocabulary* have been carefully controlled and appear to be measures of the content as taught in most basal readers. Validity is not supported by data to insure that the topics presented are actually measured on the *Reading* subtests. A careful analysis must be made by the teacher to insure that these tests match the program as taught in individual classrooms--a procedure suggested for any achievement test, particularly when data are not available. The tests are attractive, current, and both easy to administer and score. Percentiles, stanines, and grade equivalent scores are available as well as are suggestions on how to use the test results to improve a classroom program.

• Stanford Achievement Tests-Reading

Overview

The Stanford Achievement Tests–Reading are part of an achievement series designed to measure the major academic areas of the elementary and junior high curriculum. The present tests, which were published in 1964, represent the fifth revision. This review will consider only the reading subtests of the four batteries used at various grade levels in the elementary grades.

The four batteries are Primary I, used with students from the middle of first grade to the middle of second grade; Primary II, used with students from the middle of grade two to the end of grade three; Intermediate I, used with students from the beginning of grade four to the middle of grade five; and Intermediate II, used with students from the middle of grade five to the end of grade six. Each of these tests includes subtests for measuring word reading and paragraph meaning. In addition, Primary I and II and Intermediate I have a Word Study Skills subtest. Three forms (W,X,Y) are available for the Primary tests and four forms (W,X,Y,Z) are available for the Intermediate tests.

The directions for administering the tests are clear and concise and, consequently, should simplify one's administration of the tests and help to insure the test had been normed at the three different periods in the school year represented by the tables.

Each of the subtests is a timed test. The publisher suggests that the time limits ". . . are generous and calculated to give practically all pupils sufficient time to attempt all questions which the pupils are capable of answering correctly." There is no evidence given to support this statement; one may find that some of the slower readers at every grade level are unable to complete the tests. For example, on the *Word Meaning* subtest of

Intermediate II, the student is to read an incomplete sentence and select from four alternatives the correct word to complete the sentence. There are 48 of these items to be completed in 12 minutes—an average of 15 seconds per item. For the *Paragraph Meaning* subtest of the same test, the student is to read a total of 24 selections ranging in length from one sentence of only ten words to multiple-sentence paragraphs of up to 75 words. There are a total of 64 multiple choice items for these paragraphs. The total testing is 30 minutes; this time allows for an average time of slightly less than one minute for reading each selection and answering from one to five multiple choice questions.

Primary I includes three reading subtests. The *Word Reading* subtest measures the student's ability to match a picture with one of four words. Generally, the pictures do not seem to be overly biased toward a middle class population, and they are clear and easy to interpret. The *Paragraph Meaning* subtest contains 33 paragraphs with a total of 38 blanks in the paragraphs. The pupil is to supply the correct word for each blank from four alternatives. Several of the items call for the understanding of a single word. For this reason there is probably a great deal of similarity between this subtest and the *Word Meaning* subtest. Supporting this point is the fact that the correlation of the two subtests for first grade children is .72. Because of this high similarity, the two subtests should never be used as measures of distinct reading skills but should only be used as indications of general reading ability.

There are four separate parts in the *Word Study Skills* subtest. All of the tests measure the pupil's ability to match written symbols with spoken sounds. The test utilizes matching beginning sound of words and letters, matching ending sounds of words and letters, matching a spoken word with a written rhyming word, and matching a spoken word with its written form. The test correlates .73 and .67 with *Word Reading* and *Paragraph Meaning* respectively. Again, one is strongly cautioned against any attempt to utilize this score diagnostically.

For Primary II, the correlations between the *Word Reading* and *Paragraph Meaning* subtests are even higher than for Primary I. For both second and third graders the correlations are .83. Again, one is strongly cautioned against using these subtests as diagnostic measures of distinct reading skills. In fact, the test publisher should not even provide separate scores for the subtests but instead should combine them into a single reading score.

The *Word Meaning* subtest of Primary II measures the pupil's ability to pick from four alternatives the final word of an incomplete sentence. Some of the items seem to be measuring other skills than word meaning. One item tests the student's knowledge of number of items in a dozen; another item is based on whether the student knows the name of a specific country. The *Paragraph Meaning* subtest utilizes the same procedure as the Primary I test. The pupil is to supply the missing word in a paragraph. Four alternatives are supplied for each blank.

The *Word Study Skills* subtest is divided into three parts: the first two parts include auditory discrimination tests for beginning and ending

sounds and the third part measures the pupil's ability to match the underlined part of a word to a word that has the same sound; for this third part, no words are pronounced for the pupil. The correlations of this subtest with *Word Meaning* and *Paragraph Meaning* are .69 and .73 at grades two and three.

The Intermediate I subtests follow the same form as Primary II; Intermediate II is also the same, but it does not include a word study subtest. As one might expect, there are extremely high correlations between the *Word Meaning* and *Paragraph Meaning* subtests. For Intermediate I at grade 4, the correlation is .82; for Intermediate II at grades 5 and 6, the correlations are .83. Again, one must not attempt to use these separate subtests for any diagnostic purposes. The *Word Study Skills* Test for Intermediate I also correlates very highly with *Word Meaning* (.71) and *Paragraph Meaning* (.73) at grade four. Several of the items on the *Word Meaning* subtest of both Intermediate I and Intermediate II seem to be measuring knowledges other than word meanings.

Norms

The norming population for the Stanford Achievement Test is a carefully selected stratified sample from the total student population in the United States. The publisher will provide, upon request, a booklet entitled *Stanford Achievement Test: A Supplementary Report on the Norm Group.* This booklet describes in detail the relevant data regarding the norm group. If one uses the norm tables in the test manual, one will probably want to utilize these descriptions to see how a specific group compares to national population on such items as economic characteristics, regional characteristics, and size and location of community. As suggested with other tests, it would be very useful to also develop local norms. However, the norms provided by the test publisher are as representative of actual national student performance as those of any other published test available.

Form X was, however, the only form of the test standardized by the publisher. Other forms of the test were equated to this form in a study with seven school systems. Because of the lack of information regarding the correlation of these forms, there is not the same assurance that the norms for the other forms of the test are as representative of national achievement as Form X. For this reason, if only one form of the test is needed, use Form X.

Reliability

The reported reliability coefficients for the reading subtests indicate that one can be fairly certain that the score a student receives on one day will be quite similar to the score he receives on another day. These reliabilities were based only on Form X of the test and are all determined by the split-half procedure. The publisher should have reported the correlations of Form X with the other forms of the test. This information was probably available for the study in which the publisher equated the forms, but it was not reported in the technical manual. The effect of timing a test

can artificially increase the reliabilities. As most of the Stanford subtests are timed tests, the publisher should have conducted a study to determine if timing the subtests did indeed affect the reliabilities. This information was reported for the 1953 edition, but it is not reported for the 1964 edition. In general, one can be surer of the reliability of Form X than of the other forms of the test: students' performance during the tests should be carefully observed to see if there is ample time to complete the tests.

Validity

The best procedure for determining if the Stanford Reading Tests are valid measures of reading for specific purposes is to compare the content and format of the tests to the instructional program. A description of the procedures followed in developing the content outline for the test is described in the technical manual and should be carefully studied. The careful tryout and review of items by a variety of reading specialists, classroom teachers, and test developers have probably aided in the improvement of the test content. It is strongly recommended, however, that the subtests not be used for diagnostic purposes. The publishers have not developed the reading subtests for these purposes, and no diagnostic validity evidence for the subtests is presented. In fact, the intercorrelations of subtests indicate that the reading subtests all seem to be measuring the same general reading ability.

The correlations of the reading subtests with the Otis-Quick-Scoring Mental Ability Test indicate that the Otis test and the reading tests are measuring quite different abilities at the lower grade levels, but at the upper grade levels the measured abilities appear to be more similar. This result is in keeping with studies of other tests that indicate that after a student has mastered the basic skills of reading, measures of intelligence and reading are quite similar. The correlations at all grades are low enough to make valid use of both an intelligence test and the Stanford reading tests for determining need for reading improvement based on the discrepancies between reading ability and mental ability.

Summary

The Stanford Achievement Tests-Reading are carefully constructed tests for measuring general reading ability. The test norms represent an outstanding effort to develop truly representative national norms. One should find these tests quite useful in comparing one's students to national achievement levels. The subtests should not be used diagnostically.

The lack of complete data regarding the comparability of all forms leads the writer to recommend the use of Form X whenever only one form of the test is needed. This statement does not mean that one should not use the other forms, but the reliability and norming data for these forms are not so complete.

This test series is one of the better tests on the market and is found to be quite useful for measuring the general reading achievement of students.

Appendix

THE sets of charts on the following pages have been prepared as a summary of the critical points discussed more fully in the text.

The first set presents a general description of the reading readiness and the reading achievement tests. This summary includes the grade levels for which the test is intended, subtest names, and the appropriate time it takes to administer the instrument as well as the name of the test author and manufacturer.

The second set of charts is a summary evaluation of the technical evaluation of the tests as described from the technical manuals and reports provided by the publisher. A quick perusal will reveal that each test has some strengths and some alarming weaknesses. It cannot be emphasized enough that a commercial test should be carefully evaluated. Many are attractive and time saving. Most of the tests claim to measure specific reading skills. However, sufficient evidence to support the assertion that the subtests measure the skills that are inferred from the title is almost completely missing in *all* of the tests reviewed. Only the total test score and subtest scores (one or more scales) seem to be reliable enough to be used with children. Even here some tests are lacking in evidence to support their claims.

All of the reading tests reviewed in this book measure *global* reading or readiness skills. In spite of the titles, the tests are of little diagnostic value. The so-called diagnostic charts included in many of the test manuals can be misleading if used.

The major weakness of the commercial tests is also their major strength. As a global measure of reading behavior they are excellent in that they give a reliable and valid estimate of the achievement range of children in a class in comparison to a larger group. The norms for most of these tests generally are representative of national achievement; the more recent tests are greatly improved in this regard. The standardization of procedures in administering the tests are near perfect in terms of clarity of the directions provided. In addition, advanced technical techniques are being applied to most tests.

Teachers, the writer predicts, will desire more sophisticated measures as their knowledge of how a test should be used increases. Hopefully, grade level scores, short unreliable scales, and meaningless diagnostic outlines will be removed from the tests through the joint efforts of the teacher and test manufacturer.

EVALUATION OF THE TECHNICAL INFORMATION SUPPLIED BY THE PUBLISHER

Test Name	READING READINESS TESTS					ELEMENTARY LEVEL READING ACHIEVEMENT TESTS															
	Gates MacGinitie Readiness Skills Test	The Harrison-Stroud Reading Readiness Profiles	Lee-Clark Reading Readiness Test	Metropolitan Readiness Test	Murphy-Durrell Reading Readiness Analysis	California Reading Test—Lower Primary	California Reading Test—Upper Primary	Gates MacGinitie, Primary A	Gates MacGinitie, Primary B	Gates MacGinitie, Primary C	Gates MacGinitie, Primary CS	Iowa Silent Reading Tests, Survey D	Metropolitan Achievement—Primary I	Metropolitan Achievement—Primary II	Metropolitan Achievement—Elementary	Stanford Reading Tests—Primary I	Stanford Reading Tests—Primary II	Stanford Reading Tests—Intermediate I	Stanford Reading Tests—Intermediate II		
					√	*√	*√	*√								√	√	√	√	2. Evidence as given is satisfactory but not complete enough to support test purposes.	
	√	√	√*					√	√	√	√	√	√	√	√					4. Not enough information given.	
√																				5. No information given.	

Validity

Does the test measure what it purports to measure?

1. Evidence is complete and satisfactory.
2. Evidence as given is satisfactory but not complete enough to support test purposes.
3. Data given but indicate test is not valid enough for stated purposes.
4. Not enough information given.
5. No information given.

																						Reliability Are the test results consistent?
									✓	✓	✓	✓	✓		✓	✓	✓	✓				1. Evidence is complete and satisfactory.
		✓		✓	* ✓	* ✓	* ✓												✓	✓	✓	✓ 2. Evidence as given is satisfactory but not complete enough to support test purposes.
														✓								3. Data given but indicate test is not reliable enough for stated purposes.
			✓*																			4. Not enough information given.
✓	✓																					5. No information given.
																						Norms Are grade or age equivalent scores usable?
			✓																✓	✓	✓	✓ 1. Description of norming population is complete and usable.
		✓		✓	* ✓	* ✓	* ✓	✓	✓	✓	✓	✓		* ✓	* ✓	* ✓	* ✓					2. Description is not complete but norms seem usable.
	✓												✓									3. Description is complete but norms are limited for most purposes.
✓																						4. Not enough information given.
																						5. No information is given.
																						Subtests and Items
✓	✓	✓	✓	✓	✓	✓	✓						✓	✓	✓	✓	✓					1. Subtests are not long enough for reliable use (only total test scores should be used).
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2. Subtests do not seem to be valid measures of subskills (only total test scores should be used).
													✓									3. Several test items are either outdated or misleading or should be used only with special populations.

* Additional information is available from the publisher but is not part of the examiner's manual.

DESCRIPTION OF READINESS TESTS REVIEWED

Name of Test	Subtests	Publication Date	Revision Date	Authors	Publisher	Time
Gates-MacGinitie Reading Tests—Readiness Skills	Listening comprehension; auditory discrimination; visual discrimination; following directions; letter recognition; visual-motor coordination; auditory blending	1968		Gates & MacGinitie	Teachers College Press Columbia University	120 min.
Harrison-Stroud Reading Readiness Profile	Using Symbols; making visual discriminations; using the context; making auditory discriminations; using context and auditory clues; giving names of letters.	1949	1956	Harrison & Stroud	Houghton-Mifflin	80 min.
Lee-Clark Reading Readiness Test	Matching; cross-out; vocabulary and following directions; identification of letters and words.	1931	1962	Lee and Clark	California Test Bureau	20 min.
Metropolitan Readiness Tests	Word meaning; listening; matching; alphabet; numbers; copying; draw-a-man.	1933	1965	Hildreth & Griffiths	Harcourt, Brace & World	60 min.
Murphy-Durell Reading Readiness Analysis	Phonemes; letter names; learning rate	1949	1965	Murphy & Durrell	Harcourt, Brace & World	100 min.

DESCRIPTION OF ACHIEVEMENT TESTS REVIEWED

Name of Test	Subtests	Grade	Publication Date	Authors	Publisher	Time
California Reading Test	Vocabulary; comprehension					
Lower Primary	Vocabulary; comprehension	1-2	1957 (rev. 1963)	Tiegs & Clark	California Test Bur.	45 min.
Upper Primary	Vocabulary; comprehension	2.5-4.5				40 min.
Elementary		4-6				40 min.
Gates MacGinitie Reading Test	Vocabulary; comprehension					
Primary A	Vocabulary; comprehension	1	1965	Gates & MacGinitie	Teachers College Press, Columbia University	55 min.
Primary B	Vocabulary; comprehension	2				55 min.
Primary C	Speed and Accuracy	3				65 min.
Primary CS	Vocabulary; comprehension, speed and accuracy.	2,3				17 min.
Survey D		4,5,6				65 min.
Iowa Silent Reading Test: Elementary Test	Rate; comprehension; directed reading; word meaning; paragraph comprehension; sentence meaning; alphabetizing and use of index.	4-8	1943	Greene & Kelley	Harcourt, Brace & World	60 min.
Metropolitan Achievement Test						
Primary I	Word Knowledge; word discrimination, reading	H-1	1959	Bixler, et al	Harcourt, Brace & World	63 min.
Primary II	Word Knowledge; word discrimination, reading	2				65 min.
Elementary	Word Knowledge; reading	3-4				35 min.
Intermediate	Word Knowledge; reading	4-5				45 min.
Stanford Reading Test						
Primary I	Word reading; paragraph meaning		1923 (rev. 1964)	Kelley, et al	Harcourt, Brace & World	55 min.
Primary II	Word reading; paragraph meaning					60 min.
Intermediate I	Word reading; paragraph meaning					50 min.
Intermediate II	Word reading; paragraph meaning					55 min.

CURRENT TITLES IN THE READING AIDS SERIES

Conducting In-Service Programs in Reading
Aaron, Callaway, and Olson

Informal Reading Inventories
Johnson and Kress

Reading for Children Without—Our Disadvantaged Youth
Whipple and Black

Critical Reading Develops Early
Lee, Bingham, and Woelfel

Evaluating Reading and Study Skills in the Secondary Classroom
Ruth Viox

Teaching Critical Reading at the Primary Level
Stauffer and Cramer

Guidance and the Teaching of Reading
Ruth Strang

Tests of Reading Readiness and Achievement
Farr and Anastasiow

How to Read a Book
Sargent, Huus, and Andresen