DOCUMENT RESUME

EJ 291 769 TM 011 054

AUTHOR Teale, William H.; Rowley, Glenn

TITLE Standardized Testing and the Teaching of Reading: A Practical Guide with Evaluations of Reading Tests Commonly Used in Australian Schools.

PUB DATE 84

NOTE 86p.

PUB TYPE Reports - Evaluative/Feasibility (142) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS Foreign Countries; Item Analysis; *Reading Tests; *Standardized Tests; Test Norms; Test Reliability; Test Validity

IDENTIFIERS *Australia

ABSTRACT

Standardized reading tests and associated issues involved in the teaching of reading skills are discussed, with illustrative examples of reading tests commonly used in Australia. The nature of standardized reading tests is analyzed, and the circumstances under which such tests should be used are outlined. Interpretation of scores from standardized reading tests and guidelines for judging the usefulness of the tests are discussed. Tests used in the Australian community include the Australian Council for Educational Research (ACER) Primary Reading Survey, the ACER Word Identification Test, the Cooperative Reading Comprehension Test, the GAP and GAPADOL Reading Comprehension Tests, Neale Analysis of Reading Ability, Progressive Achievement Tests, Schonell Reading Tests, and Standard Reading Tests. The tests are analyzed according to the appropriateness of their items to the stated purpose of the tests, item quality, reliability and validity, norms, and convenience. For some of the instruments, passage dependence is analyzed. (TJH)

Standardised Testing and the Teaching of Reading: A Practical Guide with Evaluations of ~eading Tests Commonly Used In Australian Schools

William H. Teale

Division of Education
School of Social and Behavioral Sciences
The University of Texas at San Antonio

Glenn Rowley

School of Education
Monash University
Clayton, Victoria, Australia

1984

Table of Contents

Part 1: Standardised Reading Tests in Theory and in Practice

Part 2: Test Reviews

Introduction

Standardised reading tests have been on the educational scene for the past sixty years or so. In that time they have been used to varying degrees by schools and teachers in Australia. Over the past few years, however, the trend has been to employ them more and more frequently. Any time a teacher contemplates using or actually uses a standardised reading test, there are a number of issues which he or she should consider carefully. In Part 1 of this monograph we examine those issues in an attempt to shed light on the advantages and disadvantages of standardised testing and generally to help teachers understand better what standardised reading tests can and cannot reveal about students' reading. We end this discussion by presenting ten guidelines which can be used to evaluate particular standardised reading tests. Part 2 of the monograph consists of the actual application of these guidelines: we have reviewed eight reading tests commonly used in Australian schools. The reviews are intended to serve two purposes: (1) to provide information on the adequacy and potential usefulness of the eight tests and (2) to serve as models of how other tests which may be being used in classrooms might be assessed. Our hope is that this monograph will serve as a useful resource for teachers, administrators, and other individuals interested in reading assessment and the teaching of reading in schools.

Part 1

Standardised Reading Tests in Theory and in Practice

What are standardised reading tests?

All teachers assess their students in one way or another. This assessment may be done in an informal, perhaps unstructured way (e.g., by casual observation of children working), or by more formal procedures which we usually refer to as "tests". Even with tests, however, the degree of structure and formality varies. Teacher-made tests are generally administered with a minimum of formality, and usually their content relates very closely to the preceding teaching. On the other hand, commercially produced (often, but not always standardised) tests require a more formal situation for their administration and usually test more generalised skills which often do not relate so closely to the content which has been taught.

A standardised reading test is an instrument designed to provide a quantitative measure of one or more aspects of reading behaviour. However, it must be remembered that reading is a very complex task and that effective reading is achieved by combining a variety of abilities, processes and attitudes, not all of which are fully understood. Reading tests cannot measure every component which contributes to effective reading, and so can never present a complete picture of a child's achievements in reading. What the tests usually do is to focus on a small number of dimensions which are thought to be important, including such areas as reading comprehension, word recognition and reading vocabulary.

Reading comprehension: Typically a standardised reading test has a reading comprehension section to it. This section usually attempts to measure a student's ability to understand, on both the literal and inferential levels, the printed word as a form of communication. The method employed to measure reading comprehension may be one or more of the following:

(1) Selection followed by short-answer questions:

MY DOG

One day my dog cut his leg
on an open tin, so I put him
under my arm and ran to a shop.
Here a man wound some rag round the cut.
I then took my pet home
and made him lie down in a box of straw.

QUESTIONS TO BE ASKED

1. What did the dog cut his leg on?
2. Where was the dog taken when he cut his leg?
3. What did the man do to the dog's leg?
4. What was in the box in which the dog had to lie?

Source: The Schonell Reading Tests, Test R2. Copyright Oliver and Boyd, 1942. Reproduced by permission of the publisher.

(3) Sentence-completion:

1. Fred had five white mice. He kept them in
   a tiny hutch made of wood and -------- (a)
   One day when he went to feed the mice he
   found that they had gone. He looked around
   and found a small -------- (b) in the wire.

   (a) bread, sand, wire, leaves, paper.
   (b) pot, nut, pole, stick, hole.

Source: The Schonell Reading Tests, Test R4. Copyright 1944 by Oliver and Boyd. Reproduced by permission of the publisher.

(3) Sentence-completion:

1. Fred had five white mice. He kept them in
a tiny hutch made of wood and -------- (a)
One day when he went to feed the mice he
found that they had gone. He looked around
and found a small -------- (b) in the wire.

(a) bread, sand, wire, leaves, paper.
(b) pot, nut, pole, stick, hole.

Source: The Schonell Reading Tests, Test R4. Copyright 1944 by Oliver and Boyd. Reproduced by permission of the publisher.

(4) Fill in the blank (Cloze or modified Cloze procedures):

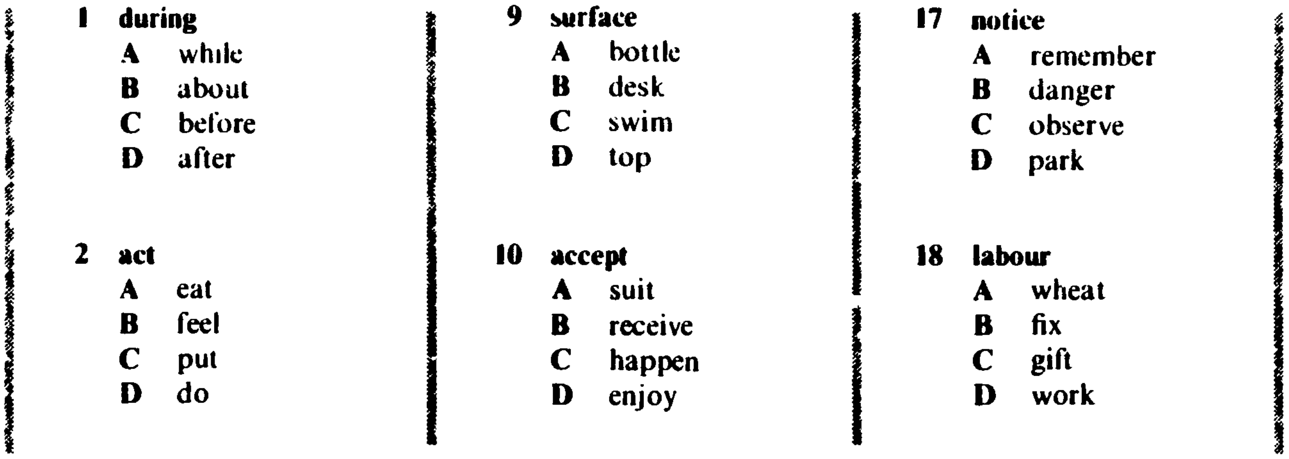
8 POLLUTED BEACHES

ANSWERS

Some of our country's beaches are
turning black. The sand __ □□□□□□□□□□
becoming dirty and oily. Even ___ □□□□□□□□□□
air around the beaches smells un-
clean. Many of our fine _______ □□□□□□□□□□
are becoming polluted with ___ . □□□□□□□□□□

Source: GAPADOL Reading Comprehension Test, Form Y. Copyright 1972 by J. McLeod and J. Anderson. Reproduced by permission of publisher Heinemann Educational Australia Pty. Ltd.

<u>Reading vocabulary and word recognition</u>: A reading vocabulary section, a word recognition section or both will also usually be found in standardised reading tests. The reading vocabulary (word knowledge) section attempts to assess the extent of the student's knowledge of word <u>meanings</u>. Here are some typical questions from the vocabulary sections of standardised reading tests:



1 during
- **A** while
- **B** about
- **C** before
- **D** after

2 act
- **A** eat
- **B** feel
- **C** put
- **D** do

9 surface
- **A** bottle
- **B** desk
- **C** swim
- **D** top

10 accept
- **A** suit
- **B** receive
- **C** happen
- **D** enjoy

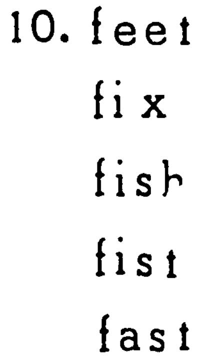
17 notice
- **A** remember
- **B** danger
- **C** observe
- **D** park

18 labour
- **A** wheat
- **B** fix
- **C** gift
- **D** work

<u>Source</u>: <u>Primary Reading Survey Tests</u>, Level B. Copyright 1976 by ACER.
Reproduced by permission of the publisher.

Source: *Progressive Achievement Tests*, Reading Vocabulary. Copyright New Zealand Council for Educational Research, 1969. Reproduced by permission of the publisher.

Word recognition refers to a person's knowledge of the correspondence between words in print and words in oral language. In the word recognitiou section of a test the student might be presented with fish and expected to say /fiš/, or the teacher might say /fiš/ and the student would be expected to pick out the appropriate graphic configura:ion from a list like this:



10. feet
fix
fish
fist
fast

Some standardised reading tests contain sub-tests in addition to the three just discussed. Frequently, there is an attempt to identify and measure component skills of word recognition, e.g., auditory discrimination, visual discrimination, blending, syllabification. Such sub-tests are generally found in diagnostic reading tests. Also some standardised tests assess speed of reading.

However, not all standardised tests have sub-tests designed to measure more than one aspect of reading behaviour. The A.C.E.R. *Word Identification Test*, fcr example, assesses only a single facet of reading, word recognition. Then, too, teachers sometimes use only one or two particular parts of a standardised test. The Schonell Reading Tests consist of four parts: a Graded Word Reading Test (R1, a word recognition test), a Simple Prose Reading Test (R2) and two Silent Reading " (R3 and R4). The R1 is frequently used by teachers in Victoria, for example, whereas the R3 and R4 are rarely administered even though they are part of the overall test. Similarly, the SRT 12 and the SRT 1 are often the only tests used from the total of twelve sub-tests which comprise Daniels and Diack's Standard Reading Tests.

	Sub-tests				
Name of Test	Reading Comp.	Vocab.	Word Rec.	Aud. Discrim.	Vis. Discrim.
A.C.E.R. Primary Reading Survey (PRS)	✓	✓	✓		
A.C.E.R. Word Identification Test (WIT)			✓		
Co-operative Reading Comprehension Tests	✓	✓			
GAP/GAPADOL	✓				
Neale Analysis of Reading Ability	✓		✓	✓	
Progressive Achievement Test (PAT)	✓	✓			
Schonell Reading Tests	✓		✓		
The Standard Reading Test (SRT)	✓		✓	✓	✓

Table 1. Content Which Selected Reading Tests Are Designed to Measure

Thus, standardised reading tests are created for the purpose of measuring one or more of the general facets of reading behaviour. In Table 1 we have provided a summary of the sub-tests contained in certain reading tests commonly used in Australia.

Generally, standardised reading tests are prepared by skilled reading and measurement professionals. The items in each test will have been extensively trialed, analysed and revised in the light of pupil responses. For these reasons, one has a right to expect that the items will be technically sound, free of the errors and ambiguities which can so easily occur in hurriedly-produced tests.

Because the content of a standardised reading test is not geared to any specific curriculum materials or reading schemes, it is intended that the test be used and re-used with a wide range of students in a variety of schools. An important property of a standardised reading test is the provision of norms. Typically, when a test is created, it is administered to a large and representative sample of children. Based on the performance of these children, tables of norms are created. Norms enable the teacher to interpret the scores of children in his/her class who have taken the test by showing how their performance compares with the performance of others in the norm group. We shall have more to say about interpreting the results from standardised reading tests further on in this booklet.

The foregoing, then, are the characteristics of standardised reading tests. We hope this discussion has helped you develop (a) a notion of the procedures by which a standardised test is created and (b) insight into what standardised reading tests actually are designed to measure.

Now that a general explanation of the nature of standardised reading tests has been provided, the next step is to ask, "Under what circumstances should such a test be used?"

Under what circumstances should a standardised reading test be used?

It should be said at the outset that although evaluation is a necessary part of teaching reading, standardised testing is not. Generally speaking, information about students' reading that a teacher would need to be able to plan effective reading instruction for those students can be obtained by ways other than standardised testing (i.e. by observation, teacher-constructed tests, informal instruments, and so forth). Thus, it is not necessary to use a standardised reading test when teaching students to read. However, the use of a standardised reading test is justifiable if (a) there is a test available which is designed to supply the typ· of information the teacher is seeking and if (b) testing is a less time consuming yet acceptable way of obtaining that information.

(a) In order to determine if there is a reading test available which is designed to supply the information sought by the teacher, two factors become important: (1) the purpose the teacher has for measuring a student's reading and (2) what a particular test is designed to measure.

(1) The purpose for which the standardised reading tests can be used

In general, there are four purposes for which standardised reading tests are used:

(i) To assess a student's achievement (attainment) in reading and thereby estimate his/her growth in reading ability.

(ii) To diagnose a student's strengths and weaknesses in reading and thereby plan instruction.

(iii) To assess success in achieving stated goals in the teaching of reading (curriculum evaluation, by individual teacher or school-wide).

(iv) To deploy resources and/or staff to school.

Notice that the first two purposes aim to help each individual student directly and that, among other things, the latter two purposes can indirectly result in benefits for students.

The primary aim of any teacher when administering a standardised reading test should be to help the individual student. Thus, a teacher would normally administer such a test for the purpose of assessing a student's reading achievement and/or diagnosing a student's reading strategies. Both of these types of information can be useful for planning reading strategies, reading abilities, reading attitudes and reading interests in order to provide the most appropriate instruction. Completing that picture is like putting together a jigsaw puzzle. Different pieces of the puzzle come from different sources. Results from a standardised reading test can supply part of the picture. For instance, knowing how Gordon performs on the A.C.E.R. <u>Primary Reading Survey</u> can give the teacher a general notion of how Gordon compares with other children in the grade. It is important to remember, however, that results from a standardised reading test, no matter how good the test is, must <u>never</u> be taken to indicate a complete representation of any student's reading. Neither any single standardised reading test nor any battery of standardised reading tests is in itself capable of providing sufficient information about a student's reading to enable the teacher to plan appropriate instruction for the student. It is always necessary to include other types of information from sources like teacher observation, informal reading inventories, miscue analysis and cumulative reading records when making instructional decisions. Standardised tests are limited in terms of what they can tell the reading teacher. However, for the teacher who knows how to use and interpret them, such tests can provide information useful for estimating growth in reading and for diagnosing reading strengths and weaknesses.

(2) What a standardised reading test is designed to measure

We discussed this point earlier and noted that, for the most part, the tests measure such things as reading comprehension, word recognition and reading vocabulary. There is also another way of looking at what standardised tests measure. Some reading tests are constructed to be attainment (achievement) tests; others are diagnostic tests. The A.C.E.R. Primary Reading Survey, the A.C.E.R. Word Identification Test, the Co-operative Reading Comprehension Test, GAP/GAPADOL, the Progressive Achievement Test and the Schonell Reading Tests are all achievement tests. They are designed to give an indication of a student's level of attainment in reading (or in some aspect of reading) vis-a-vis the students in the group upon whom the test was normed. The Neale Analysis and the Standard Reading Tests, on the other hand, are diagnostic tests. These tests are designed not only to tell the teacher the extent of the reader's achievement but also to indicate specific areas of strengths and weaknesses (e.g., comprehension, vocabulary, auditory, discrimination, syllabification, blending, and so forth).

It can be seen, then. from the preceding discussions, in (1) and (2) that assessing achievement in reading can be viewed as a way in which a standardised reading test can be used (see p. 11) and as a purpose for which the test is designed (see p. 12). So, too, does diagnosing strengths and weaknesses in reading relate both to one's purpose for using the test and what the test is designed to do. To be justified in testing children there should be a match between these two factors. That is, a teacher should know what he/she wants to accomplish by using a standardised reading test (to measure achievement, to diagnose, etc.), and the teacher should choose a test designed to fulfill that purpose. Achieving this match results in satisfying condition (a) above, i.e., the teacher has founa a test available which is designed to apply the type of information being sought.

(b) But even if criterion (a) is satisfied, testing should also be a less time-consuming yet acceptable way of obtaining the desired information in order to be justified. Teachers have more than enough to keep them busy; thus, the time factor becomes important. However, testing must still remain acceptable: for example, subjecting students to large numbers of standardised tests would not be acceptable because of its likely effects on pupils' attitudes, even though such a procedure may be thought to be parsimonious.

Thus, by ensuring that (a) there is a test available which is designed to supply the type of information the teacher is seeking and (b) testing is a less time-consuming yet acceptable way of obtaining that information, a teacher has gone a considerable way toward satisfying the conditions under which a standardised reading test should be used.

However, this is not the entire picture. Even if (a) and (b) are satisfactorily dealt with, use of a particular standardised reading test may not be justified. For instance, a teacher may want to assess a student's reading comprehension. It would certainly be possible to obtain a test designed to measure reading comprehension. Furthermore, such a test would require relatively little time to administer and would, in itself, not be "too much" testing. Therefore, the two conditions just outlined could be considered to have been satisfied. However, a crucial issue which remains is that of interpreting the results from a test. In order for a teacher to be justified in using a standardised reading test, it is essential that the teacher understands what the results mean.

Finally, a test should only be used if it is a well-constructed instrument. That is, a test should be valid, it should be reliable, and it should be adequately normed. Imagine, for example, a test designed to measure reading comprehension which assessed only literal level understandings. Such a test could not be said to be assessing reading comprehension in its fullest sense. Or, if the norms for a test were obtained from a sample of 1000 upper middle class suburban children in England, they would likely not be appropriate for many schools in Australia. Thus any teacher contemplating using a standardised reading test should look carefully at the test reliability, validity and norming sample. Otherwise, what the teacher assumes the test is measuring may not be what is measured.

The two preceding paragraphs have pointed out two more conditions which should be added: a standardised reading test should be used only if:

(c) The teacher knows how to interpret the results from the test, and

(d) The test is well-constructed and adequately normed.

These last two points are important ones and require a bit more explanation than the brief discussion above. Therefore, we have provided the following sections which examine in some depth (1) interpreting scores from standardised reading tests and (2) guidelines for evaluating the quality of standardised reading tests.

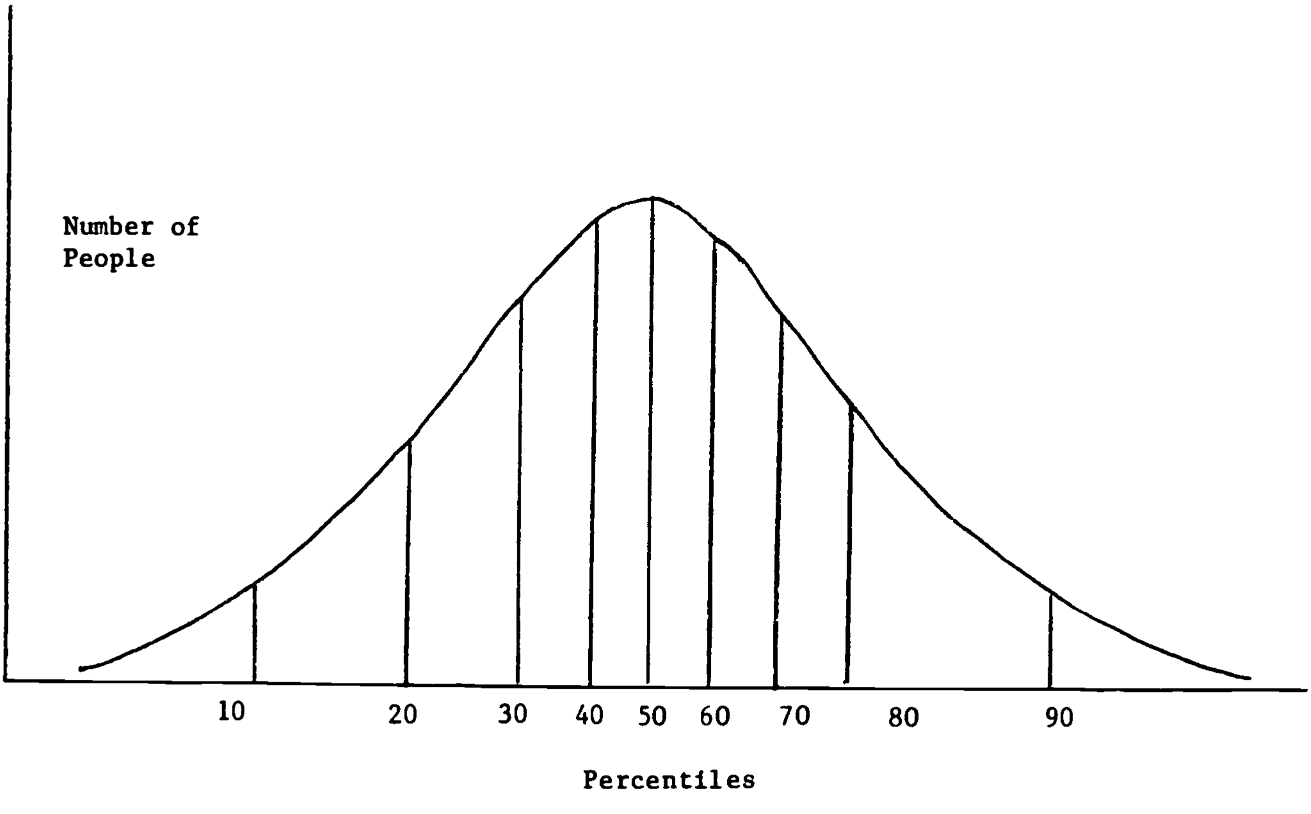
Interpreting scores from standardised tests

A raw score from a standardised reading test (the number of items answered correctly) is not very informative to most users of the test. It has therefore become customary for standardised test developers to facilitate the interpretation of performance by providing a variety of derived scores, which enable a score from one child to be interpreted in comparison with the scores of other children. Tables of norms have therefore become an accepted feature of standardised tests in all areas of achievement.

The usefulness of norms is frequently queried. The interpretation they provide is of necessity norm-referenced -- viz., an individual score is interpreted in terms of how it compares with the scores of children in the norm group. Needless to say, this will be of little value if the norm group is out-of-date, or other wise irrelevant to the situation in which the testing is done. Locating the score of a newly-arrived migrant child with little experience of speaking Engl'sh in relation to a norm group of middle-class English-speaking children may not be a very useful exercise, and would become even less helpful if the norm group were from Britain or the U.S.A. or (as is sometimes the case), if the norm data were 15 or 20 years old.

There has been, therefore, considerable pressure in recent years for the provision of aids to interpretation of test scores which are *criterion-refe:enced* -- i.e., which focus on what the child can and cannot do, rather than on how the child compares with other children. Where the pu:prse of the test is to provide instructional feedback or diagnostic information, it is easy to see why many teachers regard criterion-referenced interpretations as much more valuable. However, at the present time, the state of the art is much more advanced with respect to norm-referenced testing, and there are very few commercially-produced standardised tests available which could properly be described as criterion-referenced. 'ihe few which are available are mostly in the mathematics area, and opinions are divided about the extent to which such an approach is applicable to the evaluation of reading skills. Therefore, the discussion which follows concentrates on the interp:etation of norm-referenced standardised tests.

The aim of a norm-referenced test is to enable the user to locate a child's level of performance in relation to the range of levels of performance within the norm group. The simplest of the derived scores to calculate, and the easiest to interpret are *percentiles*. Let us say that Kathy gets a percentile of 75 on the v·.abulary section of the Progressive Achievement Test. This means that Kathy's score exceeds that of 75 per cent of the individuals in the norm group. Percentiles are straightforward to compute and easy to interpret. Perhaps their greatest advantage is that they are unlikely to be misinterpreted, since they do not carry with them any suggestions of further, hidden meanings. They have a slight disadvantage, in that they map test scores onto a scale which has unequal units. Thus the difference in achievement represented by percentiles of 85 to 95 would be much greater than the difference represented by scores of 45 and 55. The reason for this should be clear from a glance at Figure 1.



Figurel. <u>Distribution of Pecentiles</u>

Test scores, like most measurements, tend to "bunch up" in the middle ranges.

Because of this, in a group of 100 people the difference in score between the 45th and 55th persons may be only a mark or two, while the difference between the 85th and 95th persons will normally be much greater. This should not be a problem to users who are familiar with percentiles, and in fact causes no trouble at all unless we want to add, subtract, or average the scores. If such operations are carried out on percentile scores, we would want to treat the results with suspicion.

Another form of derived scores which does not have this property is the stanine (derived from the words "Standard Nines"). In converting scores in a distribution to stanines, essentially what is done is to divide the score range into nine bands of equal width (except for the two at the extremes). For the technically minded, the width of the inner seven bands is one-half of a standard deviation. The central band is allotted a score of five, and the scores range from one (the lowest) to nine (the highest). Stanines are illustrated in Figure 2.

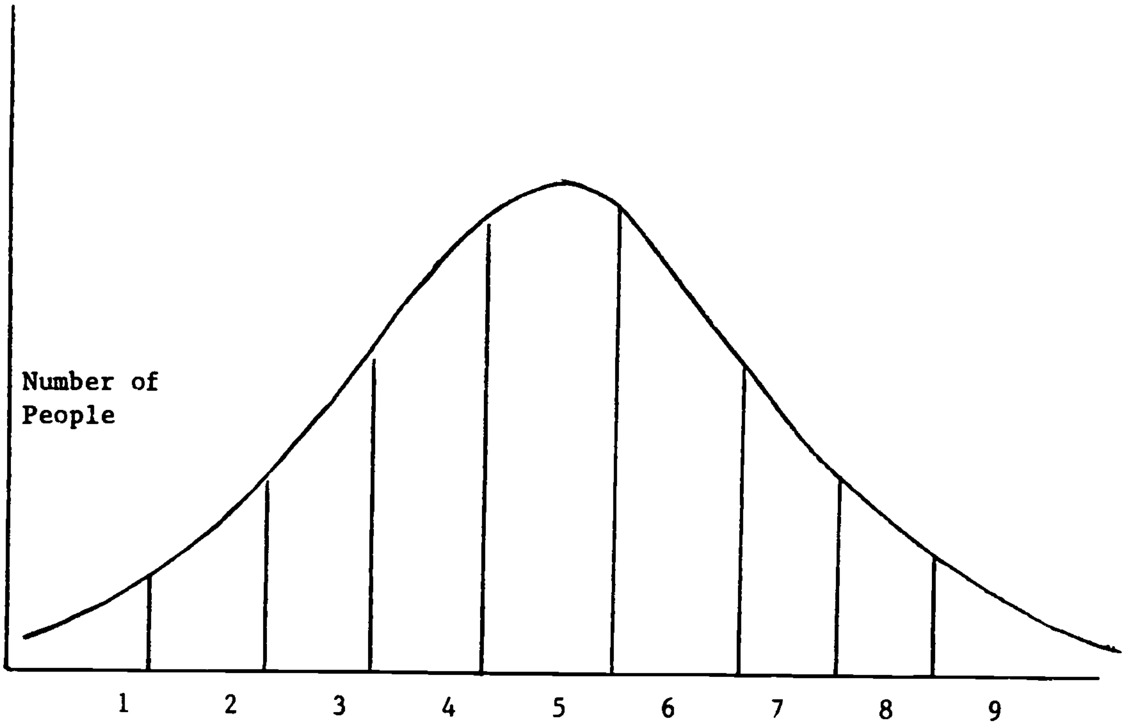


Figure 2. Distribution of Stanines

Stanines do allow for some reasonably accurate intuitive interpretations. For example:

9 = exceptionally high

8 = very high

7 = well above average

6 = slightly above average

5 = about average

4 = slightly below average

3 = well below average

2 = very low

1 = exceptionally low

Furthermore, if the score distribution is reasonably normal (and standardised tests are usually constructed to ensure that this is so), we can interpret stanines in terms of percentages of the norm group:

9 = the top 4%

8 = the next 7%

7 = the next 12%

6 = the next 17%

5 = the middle 20%

4 = the next 17%

3 = the next 12%

2 = the next 7%

1 = the lowest 4%

Thus, it is a simple matter to convert from stanines to percentiles and vice-versa, and experienced users quickly become adept at doing this.

Stanines have an advantage over percentiles insofar as interpretation of results is concerned. This advantage relates to the accuracy with which the results are expressed. All test scores contain error in the sense that they have been influenced by chance factors such as the way the child felt on the day of testing, luck with respect to guessing, or even test content or physical conditions of testing. The higher the reliability of the test, the smaller the expected errors of measurement will be. However, there are very few tests for which we could be confident in scores expressed to more than nine or ten points on a scale. Percentiles, which relate scores to a 100 point scale, can give a quite misleading picture of the accuracy of measurement involved. In simple terms, we have no reading tests which can distinguish between levels of achievement which differ by one percentile point. One stanine does represent roughly the finest discrimination we can make with reasonable confidence.

In general, if it is norm-referenced information that you seek, we can recommend percentiles and stanines as providing such information with the least likelihood of misinterpretation. Many tests, though, provide, either in place of percentiles or stanines, or in addition to them, scores which they describe as "reading ages" or "reading grades". Reading ages and reading grades are popular with teachers, administrators, and parents. We *cannot* recommend the use of reading ages or reading grades as interpretive scores. In fact, we believe that such scores are so grossly misinterpreted, with potentially unfortunate consequences, that it would be better if test publishers united in refusing to provide them. Why?

A reading age or a reading grade is a derived score which is frequently provided by reading test developers in an attempt to make test scores more readily interpretab.'e. Understandably, these test developers argue that a raw score from a standardised test (e.g., 36 items correct out of 45) is difficult to make sense of, particularly for users not thoroughly familiar with the resc. Scores such as age equivalents (of which read'ng age is just one example), and grade equivalents are provided in the belief that they add meaning to the raw scores, viz., they enable a richness of interpretation by referring raw, uninterpretable test scores to a universal and readily-understood scale (the age or grade of the child). Teachers and administrators feel comfortable in dealing with a score such as reading age or reading grade because they believe they have an intuitive feel for their meanings, a feeling which is perhaps not present when they have to deal with stanines, standard scores, percentiles, and so forth. For this reason, reading ages and, increasingly, reading grades, are commonly provided by standardised tests, and for some tests (e.g., GAP and GAPDOL) reading age is the only interpretive score which is recommended.

In spite of their widespread popu'.rity, these types of derived scores are not held in very high esteem throughout the testing profession. To understand why this is so, it will be necessary to examine briefly how reading grades are determined. Suppose we have just developed a new reading test, the Cooperative Reading Ability Profile. Prior to marketing the test, we must obtain norms. This involves our administering the test to a large and hopefully representative sample of children, similar to the children for whom the test was devised. The CRAP test was, in fact, constructed to be used with children around grades 4-6, so we go out and administer the test, to, say, 1,000 grade four, 1,000 grade five, and 1,000 grade six children. (If we wanted a reading age rather than a grade equivalent, we would choose children by age rather than by grade level.) Naturally, we will find that, children being children, there is a fairly wide range of performance within each grade level. Nevertheless, scores do tend to be higher for the higher grades, even though there is a great deal of overlap between the grades. Figure 3 probably provides a fair representation of what we might expect.

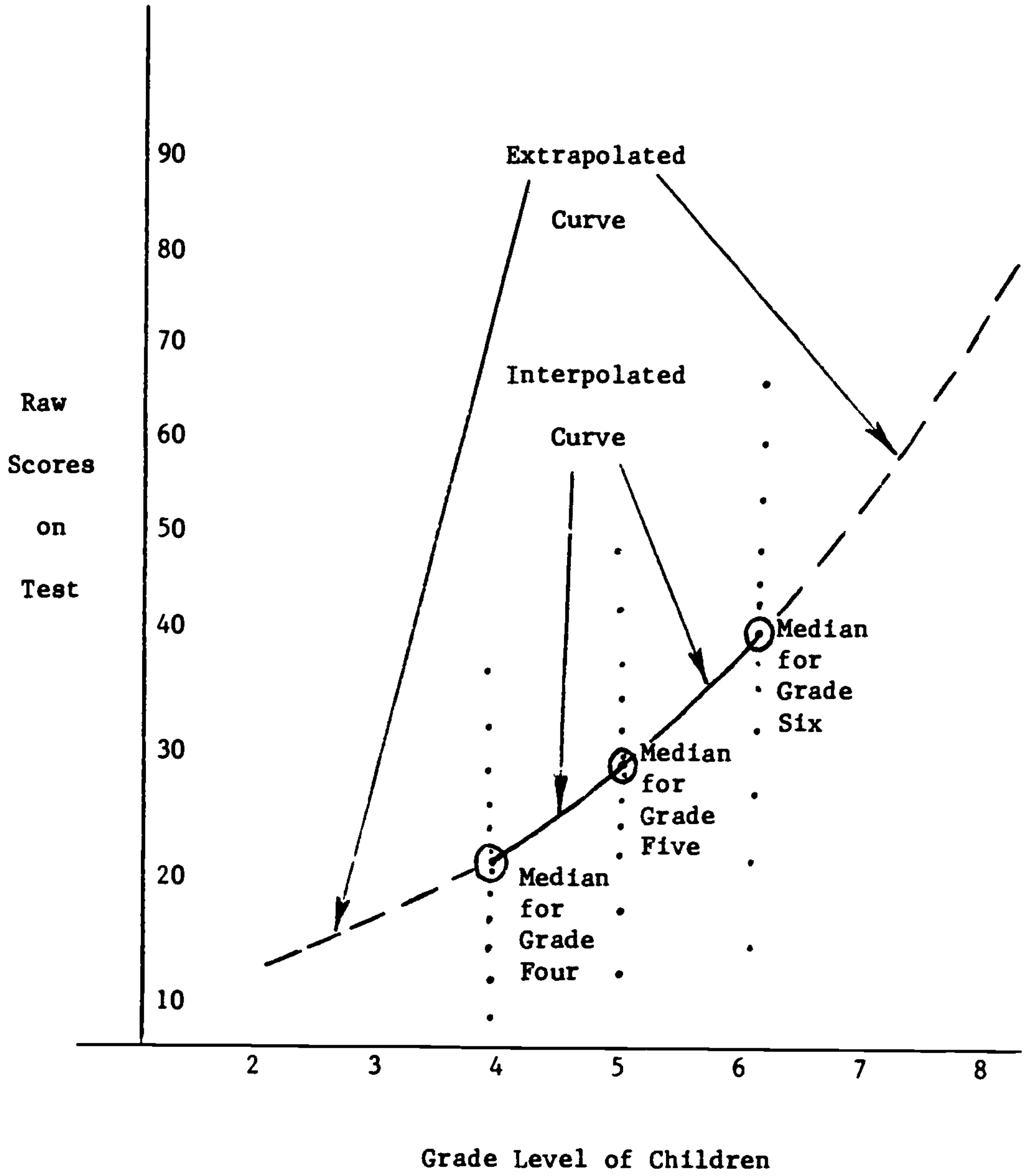


Figure 3
Curves Used to Develop Tables of Reading Grade
(Hypothetica. Test)

Thus, we note that there is no one score which describes how grade four, grade five, or grade six children perform. Instead there is a range of scores which is typical of grade fours and other (higher) ranges of scores which are typical of grade fives and grade sixes. We can, however, choose one score which is perhaps more representative of grade four ch^ildren than any other. The average score would be a defensible choice, but more commonly the one which is chosen is the median, or middle score. In Figure 3, the median scores for each grade level have been marked, and joined with a smooth unbroken curve.

Now suppose a child, at whatever grade level, comes and writes the CRAP test, obtaining a raw score of 30 points. We ncte from Figure 3 that 30 corresponds to the score of an average grade 5 child, so we allot that child a reading grade of 5.0. If a child scores 25 (halfway between a reading age of 4.0 and 5.0), we would interpolate from the graph, and allot that child a readi.g age of 4.5, or perhaps 4.6.

So far, so good. Nothing very objectionable has taken place. But you will note from Figure 3 that for many of the students in the norm group (all those with raw scores below 20 and abcve 43) we cannot provide reading grades, since there are no grades (in our sample) for which these are typical scores. The solution favoured by the test developer is to extrapolate, that is, to continue the smooth curve in either direction in the way it appears to be going. In Figure 3, the extrapolation is shown by the broken line curves.

Now we can provide reading grades for all pupils whose scores fall within the normal range. Thus, if a pupil obtains a raw score of 60, he/she will be allotted a reading grade of almost 7, and a score of 15 would gain you a reading grade of 3.5. As if by magic, we have produced reading grades of 3 and 7 without even testing children from grade three or from grade seven. But there is a price to be paid for this. Whereas we may fairly say that a reading grade of 5.0 corresponds to the score obtained by a typical fifth-grader, for reading grades in the extrapolated region, we cannot even say this. A reading grade of 7.0 corresponds to our best guess as to what a typical grade seven child might score on the CRAP test, and we have made this guess without testing children from grade seven! Have a close look at the tables in your favourite reading test which tell you how to convert raw scores to reading grades or ages. Can you tell which part of the table contains the extrapolation? If you can't, the test developers haven't been playing fair with you.

Not only do the extrapolated parts of the table represent guesses made about the performance of ficticious, untested pupils, but the chances are that they represent bad guesses. Why? You will recall that the data are obtained from a <u>sample</u> of pupils. We cannot test all the pupils of interest (e.g., all fourth-, fifth-, and sixth-graders in Australia), so we choose a sample instead. Whenever we engage in sampling, the element of chance becomes important. Sometimes samples score a little higher than the population from which they are drawn; sometimes, a little lower. This is referred to as sampling error, but it doesn't mean a mistake -- just an inevitable consequence of sampling. Suppose, in Figure 3, that the sampling error was infinitesimal for grades four and five, but amounted to one score point for grade six (i.e., the median should really have been 29, not 30). The same curve can't be used any more. Try fitting a smooth curve through the three points now. You will find that the interpolated curve is still reasonably accurate, but that the extrapolated part has shifted quite dramatically. All of this happened because of a sampling error of one score point at one grade level, and this is a margin of error which test developers would usually be pleased with. The simple facts are that extrapolated reading grades, or reading ages (a) don't mean what they appear to mean, and (b) are rather poor estimates anyway. Yet many widely used reading tests do not even provide you with the information you need in order to find out which parts of the table of reading grades are obtained by extrapolation.

There is another property of age and grade equivalent scores about which many people are unaware, although it is obvious enough if you think about it. Once a child gets to grade 6, it is possible for him/her to be two or three grades ahead (or behind) in reading, but what about in grade 1 or grade 2? Typically, scores expressed in age or grade equivalents show greatly increased variability as you proceed upwards through the grades. Thus, one year ahead of grade or age level at grade 1 may be a stupendous performance, whereas one year ahead at grade 6 may be just slightly ahead of average. Neither age nor grade equivalents tell us anything about relative performance of children from different age or gr_de levels. Moreover, we cannot accurately compare scores from one test to scores from another. Some tests produce scores which spread pupils out much more than do other tests. Thus, a nine-year-old with a reading age of ten on one test may be at a quite different level to a nine-year-old with a reading age of ten on a different test: one may be just above average and the other near to the highest in his age group. The notion, therefore, that reading age refers all scores to a universal, intuitively uaderstandable scale is a difficult one to defend.

However, the problem is not just that reading ages and similar scores are difficult to interpret. It is that they are sc easy to misinterpret. In fact, in many respects they invite misinterpretation. One particularly common misinterpretation; which is probably not confined to parents and administrators, is to view them as targets rather than as typical performances:

> "I don't know what's going on in the local primary schools. Half our Form 1 intake are performing below grade level in reading, and many of them are two or even three years behind."

Of course, it is nice that all our pupils should want to be above average and that teachers should be trying to help them, but we should not scold the teachers or the pupils too severely when we find that, despite their best efforts, half of them remain above average, and half below. This is an inevitable fact of the normal curve.

Is our example of the CRAP Test too far-fetched to be true? Some of the experiences with minimum-competency assessment in the United States suggest otherwise. A recent article in the Harvard Educational Review by Haney & Madaus (1978) provides an amusing account of the serious misunderstandings which can occur, and of the way in which they can influence policy decisions:

> "The idea of minimum-competency tests as a requirement for high school graduation was first approved by the New York State Board of Regents in March 1976. The initial plan was that tests in reading and mathematics would go into effect in 1977, and tests in civics, health and writing would begin in 1980. When ninth-grade-equivalent achievement was proposed as an appropriate minimum level to be required for high school graduation, one official in the mayor's office (apparently without any understanding of grade equivalent scores) responded, "What happened to the twelfth grade?" Board of Regents member Kenneth Clark described the New York standards as "embarrassingly too low." He said, "I am going along with it only because I believe that this is a first step, and that we can't stay too long on this first step". . . When the initially proposed tests

> were criticised as "ridiculously easy", a new set of tests was mandated by the Board to take effect in 1981. . . .
>
> In Connecticut, similar reactions were reported. When a seventh-grade-equivalent readinp level was proposed as a minimum condition for high school graduation, the school board was charged with misrepresentation: "We're paying for twelve years of schooling, but we're only getting seven years." (p. 468)

While common-sense might suggest twelfth-grade level as an appropriate minimum for high-school graduation, the inevitable result of its adoptation would be a failure rate of 50 oer cent, simply because of the norm-referenced nature of grade-equivalent scores -- a point which seems to have escaped many of the decision-makers.

A mcre subtle and more dangerous temptation is to imagine that, if a child of seven has a reading age of ten, that the seven-year-old can cope with the tasks that a ten-year-old can cope with. What such a test result really means is that the seven-year-old child can cope wit the tasks on the test about as well as the average ten-year-old can. But the average ten-year-old has had a whole set of experiences which the seven-year-old has not. So, to suggest that the seven-year-old should be expected to cope with the same work as the ten-year-old is to invite failure, disappointment and frustration.

This same problem can be seen in the use of reading grade equivalents. How do you, as a teacher, argue with the parent who pleads:

> "My daughter, who is in Grade 3, is performing at Grade 6 level in reading. Why won't they put her in a Grade 6 reader so she can work at her own level?"

Somehow you have to explain to the parent that her daughter can perhaps do Grade 3 work as well as an average grade 6 pupil can (do grade 3 work), but that this does not mean that she can do grade 6 work. In fact it is almost certain that the grade 6 children know many things which she does not know, because she has never had the opportunity to learn them. Those things do not appear on a test designed for grade 3 pupils, nor should they. But if she were to be placed in grade 6 because "that's the level she's working at", her lack of knowledge and experience in those areas could cost her dearly. We know of no other interpretations of tests scores which so invite misinterpretation, with potentially disastrous results, as age and grade equivalents.

Current opinion in the testing profession is strongly against the use of either age or grade equivalents in reporting test scores. As long ago as 1970, Lee Cronbach's standard text Essentials of Psychological Measurement commented:

> "This is as good a place as any to mention -- any condemn -- a popular but archaic conversion known as "age equivalents" and "grade equivalents" . . . In the writer's opinion, grade conversions should never be used in reporting on a pupil, or a class, or in research. . . Age conversions are also likely to be misinterpreted."

The APA Standards for Educational and Psychological Tests (American Psychological Association, 1974) recommends that "interpretive scores that lend themselves to gross misinterpretation, such as mental age or grade equivalent scores, should be abandoned, or their use discouraged." Furthermore, tests published by A.C.E.R. have not used age or grade equivalents as methods of score reporting since 1946.

Why, then, do so many developers of "modern" reading tests still defy such recommendations? Apparently they have convinced themselves that this is what teachers and administrators want, and that tests without reading ages or reading grades supplied will not sell. Perhaps they are right, but if so, it would seem to us to be, to say the least, unfortunate. Surely there is a need for us as teachers to make our views known to our colleagues, and to those who produce the tests we use.

We believe that percentiles and stanines are useful and easily understood scores which should be provided, but that reading ages or reading grades should not. Also it is important to add that a child's score on a norm-referenced standardised reading achievement test can only tell you that there _is_ a problem (if there is). It cannot tell you what the problem is, nor can it tell you what you should do about it. Only a skilled reading teaching, using a detailed knowledge of the child (which perhaps but not necessarily includes test results) can really do this. We woula ask teachers to beware particularly of test manuals which declare that pupils who get less than such-and-such a score are in need of such-and-such a treatment. There is no educational basis for such claims, and they should be treated with the scepticism we usually reserve for television commercials.

Guidelines for judging the usefulness of a standardised reading test

Although standardised tests are generally created by reading and measurement professionals, the tests available to a teacher are of variable quality. In short, some are well-constructed, and some are less than satisfactory. We offer the following guidelines as criteria by which standardised tests may be judged. The guidelines can, essentially be used as a consumer's guide to standardised reading tests. They point out what an individual can look for in judging the quality of any such test.

Guidelines	Comments
1. What does the test claim to measure?	The test manual should describe precisely what aspects of reading skill it is designed to measure. Fuzzy concepts such "reading ability" indicate lack of clarity of purpose on the test developer's part.
2. Are the items approp iate for the stated purpose of the test?	You must use your judgment here. On a test intended to measure "reading comprehension", items which could be answered merely by literal translation would not by themselves be sufficient. On a vocabulary test for primary school children one would not expect to see words which are totally outside the range of such children's experience (e.g., "ephemeral", "valetudinarian"). The issue of cultural bias is important, too. A test which assumes (or requires) certain knowledge or experience is demonstrably unfair to pupils whose background has denied them such knowledge or experience.

Guidelines	Comments
3. Are the items technically sound?	It is much easier to write inadvertently faulty items than most people imagine. Even experienced test constructors frequently fall into traps such as "giveaway" clues, items which assume special knowledge, distractors which are so implausible as to be ineffective, more than one correct (or arguably correct) alternative, and so on. Do not be overawed by the professional appearance of the test and the reputed expertise of the test developers. Item-writers are human, and fallible ones at that. It has been said, "Examinations are formidable, even to the best-prepared, for the greatest fool may ask more than the wisest man can answer".
4. On multiple choice item tests designed to measure reading comprehension, to what degree are the items "passage-dependent"? [not applicable to all tests]	Reading comprehension tests often contain a passage followed by multiple-choice items intended to assess comprehension of the material in the passage. If the items can be answered by pupils who have not read the passage, we would have to suspect that they measured something

Guidelines	Comments
	other than reading comprehension -- something perhaps better described as native wit, test-wiseness, ESP, or some combination of them.
5. Does the test manual provide sufficient evidence of validity?	Test validity has been defined as the extent to which the test "measures what it is supposed to measure". Your judgment on this matter will be made largely on the basis of a close examination of the items. Nevertheless, test developers often provide empirical evidence of validity, and this should be examined carefully.
	Validity is a topic which would require at least a chapter of a book to explain even partly satisfactorily so we will not attempt it here. [1] Validity evidence

[1] Texts in measurement and evaluation invariably contain a chapter on reliability and/or validity. Among those which we would recommend to you are Measurement and Evaluation in Teaching (N.E. Gronlund, 1976, chapters 4 and 5), Measurement and Evaluation in Psychology and Education (R.L. Thorndike & E.P. Hagen, 1977, chapter 3), and Measurement and Evaluation in Education and Psychology (W.A. Mehrens & I.J. Lehmann, 1978, chapter 5). [Originality in choosing textbook titles has not been one of the strengths of the area.]

Guidelines	Comments
	might include some or all of the following: a careful classification and description of the items in the test, measures of agreement between the test and other widely-accepted tests, eviaence that the test is predictive of children's later success in reading.
6. Does the test manual provide sufficient evidence of reliability?	A test is reliable if it measures consistently. We would have little faith in test scores if they fluctuated widely from occasion to occasion. Reliability is usually assessed by means of a measure of agreement between test and retest or between equivalent forms of the test (which is slightly better). Reliabilities estimated from the internal consistency of the test (those described as "split-half", KR20, KR21 or Cronbach's alpha) tell us little about the stability of scores, and probably indicate that test development has been done "on the cheap". One point to note: if the test is designed to produce subtest scores as well as total scores,

Guidelines

Comments

the reliability of the subtest scores is important. Frequently subtests are too short to provide reliable scores, yet teachers are urged to use them in making decisions about pupils. Although it is impossible to provide rules of thumb which are universally applicable, one should be wary of making decisions about individuals on the basis of test scores having a reliability less than 0.90, although decisions about groups (e.g., in curriculum evaluation) can be made on the basis of much lower reliabilities than this.

7. Is the test adequately normed?

What is the size and composition of the norm group? How was it chosen, and what evidence is there that it was representative of a wider population? (In this respect numbers alone are insufficient; we would need to be convinced that the sampling method provided representativeness). Are the norms relevant? (Victorian? Australian? American? Brit-

Guidelines	Comments
	ish?) How recent are they? Frequently, test manuals simply don't provide this sort of information to their users. They should.
8. Is the test convenient to use?	Teacher-made tests can be used flexibly and informally, whereas standardised tests frequently require uniform conditions of administration, a degree of formality and time allotments which may or may not fit in with the school schedule. Standardised tests are generally convenient with respect to marking and scoring, and usually the test booklets are reusable. An important consideration is the effect which the administration of the test will have upon the normal running of the class.

By using these eight guidelines to judge the quality of a standardised reading test, you can determine whether a test is well-constructed or technically unsatisfactory. But being well-constructed is not sufficient grounds for using a test. You must still consider whether the particular test will be valuable to you, the teacher. All of the circumstances outlined earlier in this booklet must be taken into account. In addition, we offer the following two guidelines which are of great importance:

Guidelines	Comments
9. How well do the (inferred) objectives measured by the test, match those being pursued by your class?	A close study of the test itself allows you to judge the abilities required for a pupil to do well on it. (These may or may not coincide with the stated purposes of the test.) How well do these abilities match those you have been promoting in your teaching? Put very simply, you must ask yourself whether the curriculum which you follow would give pupils a fair chance to do well on the test and, on the other hand, whether the results from the test can really tell you anything about how well a student has fared or will fare in your curriculum.
10. Will the test help you to teach more effectively?	This is really the ultimate consideration. Will the knowledge gained from administering the test result in more effective learning by young pupils? This may come about in a number of ways -- by enabling you to make better provision for the individual differences in

Guidelines

Comments

the class; by pointing out areas of weakness which you can remedy; by enabling you to plan a curriculum more suited to the achievement level of the class; by enabling pupils (and perhaps their parents) to understand their strengths and weaknesses better. As a teacher your responsibility is to look at the kind of information which can be obtained from the test and to judge whether that information will benefit you in your teaching and your class in their learning.

These ten guidelines are summarised on the following checklist. After close examination of a test, you may find it useful to enter your judgments on this checklist in order to see how the test stands up in terms of its quality and usefulness to you as a teacher.

CHECKLIST FOR EVALUATING A STANDARDISED READING TEST

NAME OF TEST: ______________________________

PUBLISHER: ______________________________

DATE OF PUBLICATION: ____________

CLAIMS TO MEASURE:
1. ______________________________
2. ______________________________
3. ______________________________
4. ______________________________

NORMS: YES / NO

Where: ______________

When: ______________

Sample: ______________

____

SCORES PROVIDED:

_____ Percentiles

_____ Stanines

_____ Reading Age

_____ Reading Grade

_____ Other: ____________

CHECKPOINT	RATING		
1. Stated purpose	Very clear □	Fairly clear □	Absent or unclear □
2. Appropriateness of items to stated purpose	Very Appropriate □	Fairly Appropriate □	Inappropriate □
3. Item quality	Very sound □	Mostly sound □	Many unsound □
4. Passage dependence	Most items passage dependent □	Many items passage independent □	Most items passage independent □
5. Validity evidence	Very comprehensive □	Adequate □	Inadequate □
6. Reliability evidence	Very comprehensive □	Adequate □	Inadequate □
7. Norms	Adequately described □	Inadequately described □	Not provided □
	Relevant □	Fairly relevant □	Irrelevant □
8. Convenience	Very convenient □	Fairly convenient □	Inconvenient □
9. Match with class objectives	Very close □	Fairly close □	Not close □
10. Information about child provided	Very helpful □	Fairly helpful □	Not helpful □

Conclusion

In short, we offer no easy solution to the teacher or other school personnel who contemplate using a standardised reading test. 1nere is no magic in tests that will help a person teach better. Nor can tests somehow substitute for teaching. We have outlined four conditions which should be fulfilled if one is going to use a standardised reading test: (a) The test must be designed to supply the type of information sought; (b) testing must be a less time-consuming yet acceptable way of obtaining that information; (c) the teacher and other school personnel must know how to interpret the results from the test and (d) the test should be well-constructed and adequately normed. These four criteria are rigorous ones, and so should they be. Too often standardised testing of reading has disastrous consequences for individual children because uninformed users employ the tests merely as sorting or labelling devices. The procedure of standardised testing is quite subject to abuse, and therefore we need to be very careful. It is all too easy for tests to be employed in a mindless way. A standardised reading test can provide useful information to a teacher who combines instructional techniques with the motivation to understand and help the child. But the test should never be the criterion at which instruction is aimed. The usefulness of any standardised reading test is dependent at least as much upon the teacher as it is upon the test. .hat we must always keep at the forefront of teaching is the idea that it is by responding to what the child is trying to do that a teacher can most felicitously and efficiently help that child learn to read.

Part 2

Reviews of Standardised Reading Tests Commonly Used in Australian Schools

ACER Primary Reading Survey

Publisher: Australian Council for Educational Research

197?

Stated Purpose and Description

The Primary Reading Survey (PRS) consists of a serieр of tests at six levels, as indicated in the following table:

Year Level		Name of the Test	No. of forms	Description of each form	Timing for each
1	AA	Word Recognition	2(R,S)	16 multiple-choice picture-stimulus recognition items	16 min.
2	BB	Word Knowledge	2(R,S)	20 multiple-choice synonym items	16 min
		Comprehension	2(R,S)	23 multiple-choice itemε	20 min.
3	A	Word Knowledge	1(R)	40 multiple-choice synonym items	20 min.
		Comprehension	2(R,S)	35 multiple-choice itemε	30 min.
4	B	Word Knowledge	1(R)	45 multiple-choice synonym items	20 min
		Comprehension	2(R,S)	38 multiple-choice items	30 min.
5	C	Word Knowledge	1(R)	45 multiple-choice synonym items	20 min.
		Comprehension	2(R,S)	39 multiple-choice items	30 min.

6	D	Word Knowledge	1(R)	40 multiple-choice synonym items	20 min.
		Comprehension	2(R,S)	34 multiple-choice items	30 min.

In addition, supplementary tests are provided at level D (year 6) in the "component skills" of word discrimination, word formation and dictionary skills.

Descriptions of the tests themselves follow.

Level AA (Year 1)

Word Recognition: In this test pupils are asked to look at a picture and choose from three possibilities the word closest in meaning to the picture. The sets of three words for each question, or item, have been selected from words in the low levels (1-5) of the NZCER Alphabetic Spelling List. Four types of items are represented in each form:

- The first letter or pair of letters in each of the three words is the same, but the end of each word differs. For example: note, north, noose.
- The last letter or pair of letters in each of the three choices is the same, but the beginning of each word differs. For example: hit, pit, sit.
- Beginning and ending letters or letter-clusters are the same, but central letters differ. For example: sheep, ship, shop.
- Items where the shape of the word as printed on the page varies. For example: cut, get, cup.

The rationale for this test is that these items represent common sources of reading problems experienced by beginners.

Level BB (Year 2)

Word Knowledge: In this test pupils are asked to read a word and, from the three choices next to it, choose the word closest in meaning to the key word. As in the AA test-forms, words used have been selected from the five lowest levels of the NZCER Alphabetic Spelling List. Three main types of items are used in each form:

- Closely shaded in meaning: where the distractors are close synonyms with the stem-word, but the keyed answer is closest of all.

- Associative: where distractors are associated with the stem word but not very closely, by comparison with the keyed answer.

- Partial relationship: where the keyed answer is a synonym of the stem-word, one or both of the distractors being part of the whole object or idea represented by the stem-word.

Reading Comprehension: Seven short passages of increasing size (up to 70 words) are presented, each with two or more multiple-choice questions. Pupils are asked to select the correct answer from three choices. Questions, in the main, ask the pupil to recognise a fact in the passage; one or two of the later questions will require the pupil to make an inference based on material presented in the passage. Passages present simple narratives, or descriptions and are adaptations of sections of published children's stories.

Levels A-D (Years 3-6)

At each grade level one Word Knowledge Test and two equivalent Comprehension Tests are available. The PRS is meant to be administered as a group test and the testing format is multiple choice.

Word Knowledge Test: This test is designed to assess understanding of word meanings. Students are required to choose the word closest in meaning from three or four plausible choices. Testing time: 20 minutes. Number of items: 40-45, depending upon level.

Comprehension: The comprehension tests contain a number of narrative and expository passages from a substantial variety of subject areas (Level D contains a poem and an excerpt from a play also). There are from 3 to 8 multiple choice questions for each passage. The questions are designed "to measure the pupil's comprehension of facts, inferences, implications and underlying assumptions. The distribution of literal and inferential questions is as follows:

	A	B	C	D
Literal	50%	75%	75%	75%
Inferential	50%	25%	25%	25%

Testing time: 30 minutes. Number of items: 30-40 depending upon level.

At Level D Supplementary Tests 1A, 1B and 1C are provided. There is one form of each test, and the testing format for each test is multiple-choice. The parts are as follows:

1A: *Word Discrimination* - Questions in this section require students to discriminate meanings of near synonyms. Sentence completion, selecting a synonym for a word or phrase, choosing a word or phrase, choosing a word "stronger in meaning" than a given word and selecting a noun "which, in normal usage, accompanies a particular stem adjective" are all employed in this section. 40 items. Testing time: 15 minutes.

1B: *Word Formation* - Students answer questions relating to the formation of different parts of speech by adding affixes to or otherwise changing root words. 35 items. Testing time: 30 minutes.

1C: *Dictionary Skills* - Students use a miniature dictionary of 32 words and answer questions relating to abbreviations, word meanings, word pronunciation, spelling and word formation and recognition. 30 items (4 of which are word-choice format). Testing time: 30 minutes.

Appropriateness to Stated Purpose; Item Quality

It is difficult to fault the quality of the items in these tests. In the comprehension tests the passages are interesting and appear appropriate for their intended levels. The questions are technically sound, and a nice balance is struck between literal and inferential comprehension.

In terms of "stated purpose", it should be noted that what is measured by the word knowledge tests is the ability to recognise synonyms, free of context. This presents a different (and more difficult) task than the recognition of word meanings when the words are presented *in context*, as is done, e.g., in the *PAT*. Our view is that the understanding of word meanings in context is more relevant to the task of reading and therefore more important.

For the youngest children (Level AA), the stimuli are in the form of sketches, which the child matches to the appropriate word. These seem to be mostly well-chosen, although a picture of a shovel (associated with the word "dig") is not instantly recognisable; the drawing for "pen" is an old-fashioned fountain-pen which many children will never have seen; and it seems unfortunate that the distractor "gun" in item 11 actually has a gun pointing upwards at it from item 12.

Norms

The manuals present stanines for each grade level, based on a national sample. In addition, the construction of local norms is encouraged and explained. The use of stanines is explained and justified with some care, and one must hope that this section will be read by most users. Tables of percentiles would have been useful, but the sample sizes were perhaps a little small for this.

The sampling was done with the care we have come to expect from ACER, at least for Levels A, B, C and D; the origin of the norms provided for Levels AA, BB is not adequately described. The sample size appears small (mostly less than 200 at each grade level), and one senses that budgetary problems were experienced. Nevertheless, a clearer description of how the norms were obtained should have been provided.

Reliability

The evidence on reliability is mixed, and has to be gleaned from tables at the ends of the manuals. Mostly only KR20 is provided, although with alternate forms available, the superior alternate forms coefficient could have been presented for all but Word Knowledge at Levels A-D.

The coefficients reported are acceptable, but lower than those for the PAT, which must be seen as PRS's competitors. Over Levels A to D, KR20 for comprehension ranged from .82 to .92, and for word knowledge from .76 to .89. Alternate form reliabilities for comprehension were .71, .70, .66 and .64 at Levels A, B, C, and D respectively. Inexplicably, KR20 was presented only for one form (R) of the comprehension tests; the table of statistics for form 5 was annotated "KR20...not available." This of course is a mis-statement; it's in there if someone cares to dig it out. Perhaps the manual might have read, with greater frankness, "we haven't got around to it yet."

As the lower levels (AA and BB) the KR20 coefficients were understandably lower; .67 and .75 (Word Recognitions: AA), .79 and .81 (Word Knowledge: BB), and .83 and .84 (Comprehension: BB). Again, alternate forms coefficients would have been preferable, and (one expects) lower.

In all, the information on reliability is incomplete and not overly impressive. After the mandatory section in the manual on how important reliability is, it is surprising that reliability is never mentioned again, and the relevant data are tucked away in tables as the back of the manual, and not specifically identified as reliability data.

Validity

No data are provided on validity. Users must judge from the test content whether what they measure is fitted to the user's goals. Our view is that the tests are soundly constructed, providing reasonable measures of comprehension and word knowledge bearing in mind the limitation mentioned previously--that it measures word knowledge out of context. A clear outline of test content and rationale would have been a welcome addition to the manual.

General Evaluation

Although we regard these tests as soundly built instruments with adequate psychometric properties, there are some serious concerns about them.

Firstly, the supplementary tests, which focus on the skills of word discrimination, word formation and dictionary skills, seem unsuited to the norm-referenced interpretations (stanines, percentiles) which are provided. Norm-referencing is most suited to survey tests or tests of attributes where the ranking of people in terms of "how much" makes some logical sense. For specific skills such as these, a criterion-referenced interpretation focusing on skills achieved might have been more useful.

Secondly, there is the question of difficulty. Consider the following table of mean scores, taken from the norms sections of the manuals:

Year Level		Name of the Test	Mean Score	No. of Items	Mean Score as a percentage
1	AA	Word Recognition	(R) 9.3	16	58
			(R) 9.6	16	60
2	BB	Word Knowledge	(R) 10.0	20	50
			(S) 9.3	20	47
		Comprehension	(R) 8.8	23	38
			(S) 9.8	23	43
3	A	Word Knowledge	(R) 15.9	40	40
		Comprehension	(R) 12.6	35	36
			(S) 13.3	35	38
4	B	Word Knowledge	(R) 19.2	45	43
		Comprehension	(R) 16.5	38	43
			(S) 20.0	39	51

5	C	Word Knowledge	(R)	21.6	45	48
		Comprehension	(R)	18.5	39	47
			(S)	20.0	39	51
6	D	Word Knowledge	(R)	16.9	40	42
		Comprehension	(R)	14.0	34	41
			(S)	14.3	34	42

Even at year 1 the tests are very difficult but from years 2 through 6 they are exceedingly so, with average scores around 43 percent. Our view is that there is no gain, and potentially some loss in subjecting children as young to this degree of difficulty and frustration. Remember that these are (with the exception of Levels AA and BB) four-choice items, for which the psychometrically-optimal mean score is 62.5 percent, so that the extreme difficulty does not even result in enhanced test statistics.

From where does the difficulty arise? The items in themselves do not appear all that difficult. Could it be that the tasks are too long for the time allowed? We think so. The item-statistics provided, for example, indicate that in almost all instances the last ten items on each test were answered at or below chance level (25 percent correct). If our interpretation is correct, the tests are far too highly speeded; the KR20 coefficients are grossly inflated, and, worst of all, the tests largely measure another ability which is left unstated: speed of work.

Perhaps it is a matter of priorities. Our judgment is that most teachers at these levels would prefer to encourage their students to read more slowly and carefully if that is what they need, and to be rewarded for so doing. If so, they will prefer to use tests which place less of a premium on speed of work.

For teachers who like the tests (and we emphasise that they are well-constructed) here is another recommendation: use the tests, but with more generous time allowed. Construct local norms based on the new time-limits, since the published norms will no longer be applicable. And watch carefully as the children take the tests. We do not help young children learn by setting them tasks which are beyond them.

A.C.E.R. Word Identification Test

L. Allen
Publisher: Australian Council for Educational Research

Originally published 1933
Current edition 1972

Stated Purpose and Description

The Word Identification Test (WIT) is designed "to assess the child's word identification skills and to give the teacher some insight into the child's approach to reading". The WIT is meant to indicate whether the child uses a predominantly "phonic" or "whole-word" approach with individual words. The test is to be used with children of any age who score at Year 1, 2 or 3 on a reading achievement test.

The WIT is an individually administered word recognition test containing 100 words which the child is to read to the test administrator. The test is stopped when the child "appears to be reaching either failure or frustration level".

Testing materials consist of a card on which are printed the 100 words, an information leaflet containing direction for administering the test and a sheet for recording the child's performance. This last sheet includes a table for interpreting results.

The WIT was originally one of a battery of three tests called the Individual Reading Test (1933). The words included in the WIT were selected from basal readers being used in Adelaide in 1932.

Recommended testing time: 5 to 10 minutes.

Appropriateness of Items to Stated Purpose; Item Quality

These two checkpoints are most easily discussed jointly for the Word Identification Test. In examining item quality and appropriateness of the WIT, the stated purpose of the test plays a key role. Remember that the WIT is intended to facilitate judgments about a child's approach to reading based solely upon that child's recognition of isolated words. In so doing, the test is immediately in danger of defeating its own purpose. This defect is the overriding problem with the WIT. Performance on a word recognition test is not necessarily indicative, in either a quantitative or qualitative fashion, of a child's "word identification skills" or "approach to reading" when it comes to reading connected discourse. Thus, as a test on its own rather than as a part of an overall picture of a child's reading, the Word Identification Test cannot be expected to fulfill its stated purpose.

Other shortcomings can be noted when examining the words which comprise the test. Although many of the items are appropriate for inclusion in a test of general word recognition (30 of the 100 words either appear on lists of the most frequently occurring words or are 'phonically regular'), it is obvious that a large proportion of the words have been selected because they featured in the basal reading scheme which served as the source of the words, and are quite outside the experience of today's children (e.g., "valetudinarian"). Because of the basis for item selection, the generalisability of the results from the WIT must be questioned.

Passage Dependence

Not applicable.

Reliability and Validity Evidence

No evidence of validity is provided in the technical information on the test. In addition, the validity of the WIT must be questioned on two counts. First, because the words for the test were taken from basal readers being used in Adelaide in 1932, the test would be inappropriate as an indication to the teacher of the approximate level of a child's power in word recognition in the 1980's. Teacher-designed word recognition instruments based on materials being employed in the classroom would be more appropriate if an indication of word recognition ability is being sought.

A second shortcoming of this test with respect to validity is the one alluded to earlier in the discussion of the appropriateness of the test items. Because the WIT is published as a separate instrument with its own year level interpretations, teachers may be tempted to use this test by itself as an indication of a child's reading level. Generalising about a child's overall reading level or reading strategies on the basis of results from a word recognition test will likely lead to false judgments about many children in a classroom. Remember that this test was originally intended as one of a battery of three tests. When the results from a word recognition test are looked at in conjunction with other reading abilities/strategies and relevant psychological/personal/social factors, a reasonable profile of a child's reading can be developed. However, any time a word recognition test is used as the sole indicator of a child's reading, serious misjudgments can result.

No indication of the reliability of the WIT is provided in the technical information on the test.

Norms

Results from the test are interpreted in terms of year levels. (Children who score 0-17 are equivalent to prep; children who score 36-44 are equivalent to late year 1; etc.) A.C.E.R. stresses that the norms for this test are "tentative". The grade level equivalents are based on the performance of a sample of "about 1,000" children in Victorian Education Department Schools in

1965. No indication of grade levels of the "about 1,000" children is given.

The norms are inadequately described, and provide no basis for interpreting a child's performance on the test.

Convenience

The test must be administered individually but requires only 5 to 10 minutes per pupil. Thus, it could be judged to be fairly convenient for the classroom teacher.

General Evaluation

Our recommendation is that the Word Identification Test not be used. Even as a word recognition test, the WIT'S utility must be questioned because of the validity problem inherent in the inappropriate source from which the items were obtained. Likewise, there is no assurance as to the reliability of the instrument. Moreover, to employ the A.C.E.R. Word Identification Test as an overall indicator of a child's reading achievement or word attack strategies is totally unwarranted.

In short, teachers can readily construct much more useful word recognition instruments based on reading materials normally used in the classroom.

Co-operative Reading Comprehension Test

Publisher: Australian Council for Educational Research, 1973.

Stated Purpose and Description

The Co-operative Reading Comprehension Test is a group test of reading achievement "intended primarily to help teachers...determine the general level attained by their pupils in the basic skills of reading comprehension and vocabulary use." The test contains two parts: Vocabulary and Reading Comprehension. The Vocabulary section consists of multiple-choice synonym items designed to assess knowledge of specific vocabulary. The Reading Comprehension test contains passages followed by multiple-choice items, and is designed to determine the student's ability to "understand prose", i.e., to (1) determine meaning from content, (2) organise meanings, (3) identify the writer's intended meaning, and (4) draw conclusions from the context.

There are two equivalent forms of the test, forms L and M, meant for students in grades 8-10. Each form contains 60 vocabulary items, and 60 comprehension items based on 14 prose passages. The number of items per passage varies from one to eight. There is another form, Form Y, which is intended for "late secondary and tertiary level students." Form Y contains 60 vocabulary items, and 90 reading comprehension items (the reading comprehension section contains 18 passages, with two to eight items per passage).

All forms of the test have been adapted by ACER from the American Cooperative English Test (Reading Comprehension). On Forms L and M, "some of the items were changed." On Form Y the only changes made from the U.S. edition were in spelling.

Forms L and M require 40 minutes of administration time, and yield separate scores for Vocabulary, Level of Comprehension, and Speed of Comprehension. Form Y also requires 40 minutes and yields a Vocabulary score, a Speed of Comprehension score, a Level of Comprehension score, and a Total Reading score. Form Y is infrequently used, and has not been updated since its 1964 release. Our discussion will therefore focus on Forms L and M.

Appropriateness of Items to Stated Purpose

Basically, we have few reservations on this score. The items used are orthodox in type and format, and with only occasional exceptions (see next section) seem well-chosen. In the Comprehension section, the passages have, as indicated in the manual, been chosen to be representative of a wide variety of styles and subject areas. But it is also worth noting that all of the passages used are expository in nature rather than a sample of expository and narrative. In interpreting scores from the tests, teachers might well keep in mind the limitations which this implies. The tests do not, except coincidentally, measure student's ability to read and interpret narrative text.

Item Quality

The items are technically sound, although one might wonder at the inclusion of some extremely obscure vocabulary items--e.g., how many Australian students (or teachers, for that matter) would have been exposed to the use of the word "bagatelle"? We suspect that those who have would associate it with a game played on a table, but this is not the meaning sought in the vocabulary item (Form L). Perhaps greater attention to changing patterns of language use might have resulted in more severe revisions than have taken place. Another example of an item in need of revision is item 29 (Form L), in the Reading section. In order to arrive at the correct answer, the student must, as well as reading with understanding, recognise that Massachusetts is "in the East", while New Mexico, Iowa and Nevada are not. Given that the test was "adapted" from the American version, one has to wonder how such an item surviv .

Passage Dependence

We are aware of no published research on the passage dependence of the items. Unless we are deceived by appearances, it would probably be very high, viz., a higher than chance proportion of the items could be answered without previously having read the passages.

Reliability and Validity

Statistical data presented in the manual indicate that the reliability (both test-retest and alternative-forms) is high. Evidence on validity is more difficult to interpret. A range of concurrent validity coefficients is presented for Form Y, and these are, predictably, mixed. But their relevance is difficult to estimate, given that the data are (1) old, (2) American, and (3) from a slightly different test. Our view is that validity is a question concerning the legitimacy of interpretation of test scores, which in this case is better supported by a close examination of test content than by marginally relevant correlation coefficients. A teacher who is thoroughly familiar with the nature of the tasks set by the test will likely make accurate interpretations of children's performances on the test. A teacher who has not closely studied the test content may be in danger of inferring too much. In particular, our reservations concerning the relevance of some of the items should be kept in mind.

Norms

Convenience

The test requires little administration time (40 minutes for Forms L and M, 50 minutes for Form Y), and scoring is simple. From the point of view of convenience, the tests can hardly be faulted.

Convenience

The test requires little administration time (40 minutes for Forms L and M) and scoring is simple. From the point of view of con·enience, the tests can hardly be faulted.

General Evaluation

The Co-operative Reading Comprehension Tests (Forms L and M) are well-constructed tests of vocabulary and reading comprehension which have been revised slightly (but not sufficiently) from the U.S. Co-operative English Tests. The tests

GAP and GAPADOL Reading Comprehension Tests

J. McLeod (GAP); J. McLeod & J. Anderson (GAPADOL)

Publisher: Heinemann

1965, 1967, 1977 (GAP); 1972 (GAPADOL)

Stated Purpose and Description

GAP and GAPADOL are tests of reading comprehension using a modified "cloze" technique, i.e. a sentence completion technique calling for words indicative of the testee's comprehension of the passage. Approximately every tenth word has been deleted, with the deletions being restricted to blanks for which there was a consensus from a group of "efficient readers" (undergraduate education students at the University of Queensland).

The modification to the cloze procedure is that the deletion of the words was not strictly random. The effect of the modification is that the procedure no longer can be considered a measure of the readability of the passages, but it does not mean (as some critics have suggested), that the validity of the test as a measure of reading comprehension has been lessened.

GAP was originally designed in Queensland as a screening test for readers in Grades 2 to 7 and first published in 1965, with a second edition in 1967. A British edition was published in 1965 and again in 1970. Each edition consists of a small (8 to 14 pages) manual and two forms, B & R. These are not designed as parallel forms and the author suggests that both forms be used to improve reliability. Another reason suggested for use of the two forms is so that alternative forms can be given to each child to avoid cheating. Testing time: 15-20 minutes.

GAPADOL is intended for adolescents "with a wide range of reading levels". GAPADOL also has two alternate, not parallel, forms. Each form contains six extracts. Testing time: 30 minutes.

Appropriateness of Items to Stated Purpose; Item Quality

The appropriateness of using the cloze technique as a measure of reading comprehension is a matter which has been debated, and attention will be given to this question in the section headed "Validity." Accepting for the moment that the cloze (strictly speaking, the modified cloze) technique does provide an adequate measure of reading comprehension, one still should consider whether the items chosen are suitable to the task.

Two particular faults can be identified. One is that the appropriateness of the items may be lessened by cultural bias. An example is one item in the GAP test which is based on a passage from "Mary Poppins." Students familiar with the story (as many, perhaps most children would have been at the time the test was produced) would have known without reading the passage how Mary Poppins flew. Students unfamiliar with "Mary Poppins" (perhaps mostly non-Anglo-Saxon children) have a difficult task in answering the question, since they have to read well beyond the point at which the gap appears before they find any clue to the missing word. This could have implications not only for the item concerned, but for the time students have to complete tne remaining items, and for their frame of mind as they attack later items. On other occasions, specific information is called for which is not contained in the passage. In GAPADOL, for example, a passage on turtles requires students to speculate upon how much meat can be taken from a grown turtle (over 120 pounds, or over 120 kilograms?). Another can be answered from the knowledge that parts of the turtle are used to make turtle soup. In neither case does the passage provide information--the student has to provide the most plausible word, and this is a much easier task if general knowledge enables you to answer with certainty.

The modification to the cloze procedure should have avoided these kinds of situations, and to the extent that it did, it will have improved the test. That it did not reflects the nature of the group from whom consensus was sought (education students from the University of Queensland). Not only would this group be efficient readers, as the test authors intended, but also they could be expected to be better-informed and more widely-read than than the students for whom the test was intended. Thus consensus can come about not only because of redundancy in the passage (as intended), but also because of specific knowledge possessed by the University students.

Norms

GAP was originally normed on a sample of 2029 students from a upper middle class suburb of Brisbane. Age levels for this sample are not mentioned, but it appears that scores at each end of the range are extrapolated. In fact, the sample is not clearly defined and cannot be considered as representative of a general population. It is stated by the author that the 3rd edition was modified and restandardised in U.S.A. and U.K., but details of the renorming are not provided.

GAPADOL was originally normed in the New England area of northern New South Wales in the 1960's, and it was later renormed in Canada in the 1970's by J. McLeod.

It is problematical to what population either set of norms are generalisable, and the GAP test was severely criticised in a Mental Measurements Yearbook review on this account. But perhaps more to the point is the question "What did the norming process yield?" The answer, unfortunately, is (for the GAP test) just reading ages. The use of reading ages has been condemned by measurement experts for many years now, and percentile norms are generally agreed to be much more meaningful, provided that the norm group are truly representative. It is unfortunate that nothing is provided except reading ages, and there are reasons to be sceptical about them--particularly in the upper (extrapolated) regions of the tables, where the increment in reading age, which previously had varied around 1 to 3 months per correct answer, suddenly jumps to seven months per correct answer.

In addition to reading ages, tables of "Quotients" are provided, but as these are merely derived from reading age and chronological age, they can be no more meaningful, and in view of the commonly observed nonlinear relationship between reading age and chronological age, probably even less so.

Reliability

Reliability figures reported in the third edition of GAP range from .90 to .94. However, these figures were arrived at by use of split-half technique and thus must be viewed with scepticism. Furthermore, no indication of the numbers or types of subjects involved in the reliability studies is given.

For GAPADOL the manual reports internal consistency coefficients at five year levels ranging from .84 to .93. Again no indication is given of numbers of subjects involved, and, furthermore, for GAPADOL there is insufficient information about how these coefficients were arrived at.

Concerning reliability, it can be said that such information as is available indicates that both GAP and GAPADOL are acceptably reliable, but that insufficient details have been provided as to the manner in which the information was obtained.

Validity

It is in the area of validity that some of the more serious questions about GAP and GAPADOL arise. Certainly, it is difficult to argue that the tests have face validity, since the nature of the tasks given to students is quite different to that commonly associated with reading comprehension. Filling in gaps in prose material requires certain skills which may be correlated with the ability to comprehend the material, but can not necessarily be equated with it--for example, the recognition of familiar grammatical patterns. Evidence for the validity of the procedure is therefore based exclusively on correlations with other measures of reading comprehension, and one has to go beyond the manuals to find it.

There is no information on validity in the GAP or GAPADOL manuals. In his review of the GAP test in Mental Measurements Yearbook VI, E. F. Rankin does cite concurrent validity coefficients contained in the manual for the British edition. Correlations with reading tests of Schonell and Watts ranged from .75 to .81, which is certainly acceptable, and as high as the correlations of the tests with one another. Put another way, the GAP test has as much in common with other reading tests as they have with one another. This does not mean, as one could mistakenly infer, that it measures the same thing. It measures something different because the nature of the tasks presented is different. By analogy, the performance of long distance runners is highly correlated with lung and heart capacity, but when we measure lung and heart capacity, we do not imagine that what we are measuring is long-distance-running ability.

Questions of validity are questions about the appropriateness of inferences made from test scores. When a reading comprehension test requires students to read a passage, and then to demonstrate their comprehension of the passage by answering questions about the meaning of the passage, we can feel confident that high scorers have truly demonstrated comprehension, at least if scrutiny of the items reveals no other way in which the correct answers could have been arrived at. But with the cloze procedure, there are usually other ways to arrive at the correct answer, and therefore evidence of concurrent validity, while desirable, can never be quite sufficient.

Every child must attempt every passage in each of the tests. As the passages very from easy to hard, many children may experience considerable frustration in taking the test. On the GAP test a raw score of 2 out of 42 is average for a 7-3 year old child. Thus, the validity of the test for younger children must seriously be questioned. It appears that GAP has attempted to cover too wide an age range.

General Evaluation

Some theoretical questions about the modification to the cloze procedure have been aimed at GAP and GAPADOL, but these miss the point. Certainly the effect of the modification is to invalidate the tests as measures of readability or redundancy of the passages, but it is quite likely that they would increase the validity of the measures as measures of reading comprehension. Carried to its logical extreme, drastic modification of the cloze procedure to the extent that the words omitted were all words which could be deduced or inferred from a thorough understanding of the meaning of the passage, would make the test even more valid as a measure of reading comprehension, but even less valid as a measure of the readability of the passage. As Rankin pointed out in his MMY review, the changes would be justified if they resulted in a more refined measure, but there is no evidence that this is so.

Two considerations seem more to the point. Firstly, would the tests be better than a teacher-constructed cloze (or modified cloze) test, using materials chosen because they are closer to the interests of the students? The evidence seems to suggest that the answer is "no". Rankin mentioned validity coefficients for teacher-made tests which were just as high as for the GAP test, and concluded that "GAP is not superior to a cloze test that any classroom teacher could devise."

Secondly, compared to more orthodox reading-comprehension tests, GAP and GAPDOL suffer from difficulties of interpretation in that if a student performs badly on the test, it is not clear what the student failed to do. What do you teach a child who is not so good at "filling in the gaps"?

A possible reason for using these tests would be the provision of reliable, easily-interpreted national or local norms, but the norms tables provided do not inspire this degree of confidence. There are indeed severe problems with the norming procedures used with GAP and GAPADOL, and thus the possible advantage of being able to use norms to interpret results is largely vitiated. Results just as satisfactory can be obtained by using teacher-produced cloze tests together with criteria for interpretation put forward by Bormuth (1975) in his article "Literacy in the Classroom".

GAP and GAPADOL provide stimulating and interesting tasks for studeuts at the appropriate levels. B'ıt a classroom teacher with a sound understanding of his or her children's abilities and interests could produce tests which are just as sound technically, and more attuned to the interests of the children.

Neale Analy3is of Readiug Ability

Marie D. Neale

Publisher: Macmillan Education

1958, 1966

Stated Purpose and Description

The The Neale Analysis of Reading Ability is a diagnostic reading test which is intended to assess a child's achievement, weaknesses, types of error, persistence in, and attitudes towards reading. The test was devised in order to "...bridge the gap between those who rely largely on observation and personal judgment in assessing a child's ability to read and those who accept the limited information of quantitative scores on standardized tests."

The Neale Analysis consists of a test booklet containing three forms of the test, a manual of directions and norms, and record sheets. Each form of the test consists of six narrative passages of graded difficulty and increased length. Each passage is illustrated by a picture to "set the scene." The child is required to read each passage aloud, and comprehension is measured entirely through recall. The test administrator is asked to code the child's errors into the following categories: mispronunciations, substitutions, additions, omissions and reversals. Following the reading, the child is asked from four to eight comprehension questions on the content of each passage.

Three supplementary diagnostic tests--(1) recognition of letters and sounds, 2) auditory discrimination through spelling, and blending and recognition of syllables are also included in the Neale Analysis. These three tests are not designed to be scored; rather they are intended for "noting particular difficulties of an individual child with basic processes in reading."

The test is administered individually, and requires ten to fifteen minutes to complete.

Appropriateness of Items to Stated Purpose; Item Quality

The passages themselves are suitable for their intended purpose, but for modern Australian children, they have the appearance of being dated, and of being oriented more to the experiences of British children. The comprehension questions are well-written, and include interpretative as well as strictly literal questions.

The diagnostic function of the test is really quite limited, in that it produces a categorisation of the child's errors but nothing in the way of guidelines for the interpretation of those errors, or suggestions as to what could be done with the information. What is the teacher to do with the finding that Johnny makes an excessive number of mispronunciations, or an average number of reversals? In the medical sense, these are symptoms, not causes, and a medical analysis which stopped at this point would not be thought of as a diagnosis. It may be that experienced clinicians can obtain useful information from the tests, but teachers not versed in diagnosis would do well if they gained any more than the quantitative information typically obtained from standardised reading tests.

Although no details are provided, the passages are intended to be ordered in terms of word difficulty, sentence structure, and optimal length. Reviewers have noted an apparent discrepancy in that the sixth passage in each form seems more difficult than those which follow. Data provided in the manual might have given an opportunity to evaluate this criticism, but are absent.

Passage Dependence

The comprehension questions are in short-answer format, minimising the likelihood of passage independence. Although no research studies are known, it would appear not to be a problem with this test.

Reliability and Validity

Scores are provided for oral reading rate, accuracy and comprehension. Alternative-forms reliabilities cited for accuracy and comprehension are really very high (.92 to .98, within year groups). No reliability figures are provided for the measurement of reading rate, but this would be of relatively minor importance to most reading teachers.

Validity data provided included factor analysis and correlational studies. It is difficult to know what to make of the validity data. High correlations have been found between pooled scores on the Neale tests and pooled scores on the Ballard One-Minute Reading Test, and with tests of comprehension and word-recognition. Why the Neale test scores were pooled is puzzling, since this is not the way the scores would normally be treated. It would have made more sense to present the correlations associated with the separate accuracy and comprehension scores. Given the high correlations obtained with the pooled scores, one can only surmise that the correlations with the separate scores, had they been presented, would have been high.

With regard to validity, it is important to remember that a test is not valid in itself. Validity is a property of particular interpretations of test scores. Thus the validity data presented throw light on the interpretation of the accuracy and comprehension scores, but they have no relevance to the diagnostic function of the test, which is seen as its main function. Are the categories of error the most useful ones which could have been chosen? Are they usable and unambiguous in the hands of experienced, and inexperienced users? Does a profile of errors enable a teacher to prescribe appropriate activities which result in greater improvement in reading than would otherwise have been expected? These are the kinds of questions which are implied by considerations of the validity of the tests as diagnostic instruments, and they cannot be answered easily. But these are the questions which future research on the Neale Tests might usefully ask.

Norms

Reading ages are the only form of normative data provided. The problems with reading ages have been noted earlier in our discussion, and it is unfortunate that no other interpretive scores are provided. Reading ages, it may be noted, can be computed with relatively small numbers of subjects at each age level, since only the median has to be computed. Stanines and percentiles, universally regarded as more informative, are also more expensive, since they require larger numbers at each level to provide acceptable accuracy. It has been suggested that the standardisation sample was too small to allow anything but reading ages to be computed.

The standardisation sample consisted of 2262 children in England (1,221 took Form A; 552 took Form B; 489 took Form C), spread over seven year groups. Clearly percentile norms within age groups, particularly for Form B and C, would have suffered from the small sample size. The representativeness of the sample is impossible to judge. The author states that size of school, geographical area, social background, age and sex have been controlled, but provides no details, so it is difficult to know to what group the norming tables relate. No Australian norms are available. Since the standardisation sample had no children older than eleven, the higher reading ages inferred (up to thirteen) have been extrapolated and should be treated with the utmost caution.

Convenience

The directions for the test are reasonably straightforward, and the test itself is easy to administer. The coding of errors is a fairly difficult task, and expertise would come only with considerable experience. The record sheets can be faulted on several counts. There is no space for recording answers to comprehension questions, and the space provided for coding errors is insufficient. Revision of the record sheets could help administrators (particularly inexperienced ones) considerably. Because all of the scoring takes place during testing, the administrator needs to be very familiar with the test beforehand.

General Evaluation

Reviewer's opinions about the test have been varied. While it is better constructed than most, the British orientation is obvious, and the only norms available are from British children, and the representativeness of the sample is doubtful. The use of reading age is unfortunate.

Nevertheless, many of the passages are well-chosen, and the accompanying questions are generally well-constructed, so that much useful information about a child's reading could be obtained from the use of the test materials. However the problem lies in the interpretation of the information so obtained (i.e., understanding the patterns of miscues which are coded, and drawing appropriate inferences about the child's comprehension). The manual really does not provide very much help in this. Certainly a skilled diagnostician could gain valuable insights into the child's reading, but such a person could probably do as well without the Neale tests, using informal procedures and with reading materials more suited to the interests of the individual child.

The three supplementary diagnostic tests are not as useful as they might have been. For the classroom teacher, instructions for their use and for the interpretation of performances seem essential. Much work in the area of validation also needs to be done before these tests can be regarded as satisfactory instruments for diagnosis.

In short, the <u>Neale Analysis</u> has not really succeeded in "bridging the gap" as was originally intended. It is most useful to the experienced clinician, but such a person would probably have access to instruments and techniques which draw more directly upon the experiences and interests of present-day Australian children. For the classroom teacher, the test will not provide the kind of information which can be readily translated into an effective teaching strategy. A teacher, however, who is willing to make a close study of children's responses to the test, may informally arrive at an enhanced understanding of some of their strengths and weaknesses. Used in this way, rather than as a test, and particularly if rewritten to suit Australian conditions, the instrument is not without potential usefulness.

Progressive Achievement Tests

W.B. Elley & N. Reid

Publisher: New Zealand Council for Educational Research, 1969
Australian Edition: Australian Council for Educational Research, 1973

Stated Purpose and Description

The Progressive Achievement Tests (PAT) claim to measure (1) "general level of word knowledge" (reading vocabulary) and (2) "Comprehension and interpretation of prose material". The PAT are reading achievement tests designed for students in grades three through nine (ages 8-14). There are two equivalent reading vocabulary tests (Forms A and B) and two equivalent reading comprehension tests (Forms A and B) for each grade level. Either the vocabulary or comprehension test can be given by itself, but assessment with both is recommended. The PAT are group tests and have a multiple-choice format.

Reading Vocabulary Test: This test is designed to measure word knowledge. Each word is placed in the context of a short sentence, and the pupil's task is to select the best synonym from five plausible alternatives. Number of items: 45-65, depending upon level. Testing time: 30 minutes.

Reading Comprehension Test: This test is designed to measure factual and inferential comprehension of narrative, descriptive and expository prose. Each passage is from 200 to 300 words in length and is followed by 4-6 multiple-choice items, each item having five alternatives. The proportion of expository passages and of inferential items are increased for older students. Number of items: 40-47. depending upon level. Testing time: 40 minutes.

Appropriateness of Items to Stated Purpose; Item Quality

The vocabulary items are unexceptional, as vocabulary items ought to be. The greater difficulty needed for the later sections of the test has not been achieved by the rise of exotic or outdated words; indeed all words were selected from a book of common English words. In addition the word is presented as part of a sentence, so that the student meets it in context rather than in isolation.

The technical quality of the comprehension items is difficult to fault, but some doubts arise concerning the use of the New Zealand developed items (with "minor editing") in Australian schools. For example, a passage on koala bears (sic) is followed by the question "What colour are most koala bears?" It may be that such an item tests reading comprehension in New Zealand, but one would think Australian children could answer it from general knowledge. However, such examples are rare. With careful revision rather than minor editing, they might have been non-existent.

Careful control has been exerted over the balance between factual and inferential items, with the balance shifting from factual to inferential as you move upwards through the levels. Given that they correlate very highly anyway, this probably has little impact on the statistical properties of the test, but seems justified in terms of the suitability of the tasks to the children.

In the manual, great care is taken in explaining the basis upon which the tests were developed, and a user who has read this part of the manual will have a clear picture of the intended purpose of the test, for which the items are well-suited. A later section of the manual lists 12 uses to which the tests may be put, ranging from the easily acceptable ("Identifying students especially advanced or retarded in basic reading comprehension and/or word knowledge...") to the arguables ("comparing a student's reading abilities to his achievements on other areas"). In fact very few of the 12 suggested uses are supported by the validity evidence cited, and perhaps a statement to this effect, or a more modest statement of purpose, would have been appropriate.

Norms

The tests were originally normed in New Zealand, but for publication in Australia they were renormed in 1970. The Australian sample appears to have been selected with great care, and is probably about as representative as one could hope to achieve. (Mention is made of refusals, but there is no indication whether the number of schools refusing was high or low).

The interpretive scores provided are percentile ranks and stanines for each grade level, separately by state. These are the most useful and meaningful norms to provide, and ACER should be commended for refraining from providing the more salable but misleading age or grade equivalent scores.

A useful feature initiated by NZCER is a "Level Score Scale Supplement" which relates scores on the tests to reading levels defined by readability ratings, and by illustrative reading materials of suitable difficulty. This, in fact, provides a criterion-referenced interpretation for test scores which appears to be a significant advance, if it can be soundly validated by research. The tables provided are based on New Zealand data, and an Australian-based scheme seems to be most desirable as a next step.

Reliability and Validity

As would be expected of a test with such technically skilled preparation, the reliability data indicate that all is well. From the New Zealand standardisation, split-half reliabilities ranged from .91 to .94, and equivalent form reliabilities from .85 to .94, all, of course, within grades. From the Australian sample, only KR20 (internal consistency) coefficients were reported, and they also ranged from .91 to .94 (New South Wales only). It was disappointing that the preferable, but more troublesome, equivalent forms reliability was not obtained for the Australian sample. The authors seemed to be aware of this, as they claimed almost apologetically that KR20 "is a measure of internal consistency,... and is, in general, a good estimation of test-retest reliability."

Reams of statistical data are provided which relate to the validity of the tests. In brief they show that these tests correlate pretty well with similar tests, but mostly they correlate best of all with one another. Amid the myriad correlation coefficients provided (21 from New Zealand, and an incredible 54 from New South wales) the only jarring note is provided by correlations of .86 and .87 with Cooperative Reading Test M: Speed of Comprehension. One would not like to think that the test is overly speeded, and it is encouraging that these high correlations (which occurred with Comprehension Form B) were not supported by Form A or by either form of the Vocabulary Test.

Most users will rightly be more concerned with issues of content validity, and the care taken in designing the tests, along with the wealth of detail provided in the Manual, go a long way to ensuring that scores will be accurately interpreted by those users who have familiarised themselves with the tests and the early parts (pp. 1-5, particularly) of the Manual.

A concern is the fairly strong New Zealand flavour in the passages used in the Comprehension section. Whether this will make the tests less interesting and motivating to Australian children is problematical. Evidence can be gleaned from the Manual which suggests that it does not, as might have been expected, have an impact on difficulty. Items 10-50, which are answered by grade 4 pupils, make a comparison possible. On Form A, none of the eight passages are identifiably set in New Zealand. On Form B, four of the eight passages are clearly set in New Zealand, three of them involving Maori words which would be unfamiliar to most Australians (kakas, tuis, ngaio, weka). Did the New Zealanders gain an advantage on Form B? It appears not. Combining data from Tables T1 and T9, we find that, when the six Australian states plus ACT and New Zealand are ranked in order of mean score, New Zealand ranks fourth on Form A and fifth on Form B. The New Zealand fourth-graders were if anything, a little worse in relative standing on Form B.

However, even if the New Zealand flavour has not added to the difficulty of the test for Australian students, teachers may prefer passages of more local interest on the grounds that they relate more to children's normal reading. This is a legitimate concern which the publishers of the test (ACER) might consider.

General Evaluation

One of the strengths of the PAT is the Teachers Handbook, which is extremely comprehensive, and provides, aside from a wealth of normative data some informative sections on the interpretation of test scores and sensible cautions concerning their accuracy. Teachers who study the material in the Manual should avoid most of the errors which have caused some to doubt the value of any standardised tests.

Schonell Reading Tests

F. J. Schonell

Publisher: Oliver & Boyd

1942-1945

Stated Purpose and Description

The original battery of Schonell Reading and Diagnostic Tests which appeared in the author's Backwardness in the Basic Subjects (Schonell, 1942) consisted of seven tests. Currently the first four of these are being published and used in Australia, and only these four will be reviewed.

The Graded Word Reading Test (R1; 1945) is a test for determining "the level a pupil has reached in his power of word recognition." It is to be administered for the purpose of "estimating the level reached...in the mechanics of reading." The test consists of 100 words which are graded in difficulty from age 5 to age 15 (there are 10 words per year from ages 5 to 13 and 10 for ages 14 and 15 combined). The test is administered individually and the child is required to provided the correct oral representation for the graphic stimulus. There is no time limit, but about five or six minutes per examinee is typical.

The Simple Prose Reading Test (R2; 1942) is an individually-administered test in which the child reads aloud a story about a dog. Following the story are 15 literal-level comprehension questions which the teacher asks orally. The test is offered as a reading comprehension and word recognition test, with the option of also measuring reading speed.

The Silent Reading Tests A and B (R3 and R4; 1944) consist of short paragraphs followed by questions to be answered: in the case of R3 the questions are literal-comprehension questions to which the child supplies an answer, while R4 consists of multiple-choice question in which the child supplies a word which is omitted from the paragraph. It is claimed that the tests measure silent reading comprehension, and they can be administered as either group or individual tests.

A Handbook of Instructions is available, containing instructions for administration, but for detailed information and all technical data, the reader is referred to Backwardness in the Basic Subjects.

Appropriateness of Items to Stated Purpose; Item Quality

For these tests, it is difficult to see how these two issues can be separated, and they will therefore be discussed together.

R1: The 100 words were selected from a sample of 300 administered to approximately 60 children in each age group. The arrangement of items in order of difficulty is based on these testings, the easiest being read correctly by 55 percent of Schonell's five-year-olds, and the most difficult being read correctly by 45 percent of his 14/15 year-old sample. Perhaps surprisingly, given the age of the test, very few of the words (especially in the easier sections) appear likely to have fallen into disuse. However in the later sections, the search for words of sufficient difficulty appears to have resulted only rarely in words which are widely-used but misspelled (e.g., statistics), and much more frequently in words which are difficult largely because of the rarity of their use (e.g., judicature, sonambulist). As a result, a high scorer (especially in the 13-15 year levels) is a person who can recognise and correctly pronounce very unfamiliar words, and not necessarily a person who can read common but difficult words flawlessly. With this reservation, the items do seem appropriate to the stated purposes of the test. Given changing emphases in the teaching of reading, these purposes may not be as vital to teachers as they were in 1945.

R2: The questions relate very specifically to the information in the passage, so that the task presented to the child has two elements to it--to understand what is being asked, and to find the answer in the passage. The answer is always contained in the passage. This is comprehension at a very literal level, and the results of the tests would need to be interpreted with this limitation in mind. The term "reading comprehension" in the stated purpose of the test would be taken by many to mean a great deal more than being able to answer literal level comprehension questions.

R3 and R4: There are several different types of items in these two tests. Many are literal comprehension questions: read the question and look in the passage for the answer. Others require the answer to be translated into a synonym before it is selected (e.g., in a passage about "tightly-shut windows", the child has to select "closed"). Others require inference drawing upon experience, e.g., "I can skip, I go to school every day, I wear a pretty dress, I have long hair. What am I?". (One cannot help but wonder how this item would be answered by children in sections of San Francisco.) In others there is no real basis for answering except that one response is more plausible than the alternatives (e.g., Fred kept mice in a hutch made of--bread/sand/wire/leaves/paper). Others present fairly mechanical tasks, such as counting the number of letters in a word, where the only comprehension involved is comprehension of the nature of the task. It is not necessarily wrong to include such a variety of tasks in a single test if the tasks are all seen as important. But to sum all the items together and believe that they measure a single attribute called "reading comprehension" is to stretch one' credulity too far. No rationale is provided for the selection of tasks, no table of specifications, no domain-definition. There is little basis for inferring what the tests measure, except that each of the tasks involves reading skills of one kind or another.

Passage Dependence

Although we have no research evidence concerning passage dependence, there is reason to think that some of the items in R3 may be passage-independent. For example, item 1 asks, "Where is the bird's home"?, and the answer is "In a tree." Item 6, on traffic lights, asks which light is used for "get ready." And item 13, after a paragraph explaining that Christmas Day in Australia occurs in the summer asks whether Christmas Day in Australia is likely to be windy, freezing, hot, cold, or frosty. One would imagine that Australian children would not need the passage ir 'rder to answer this question, and that English children with even a modicum of testwiseness would not choose the second, fourth or fifth alternatives on the grounds that each implies the correctness of the others.

Passage dependence is nct an issue with R1, R2 and R4 because of their format.

Reliability and Validity

No reliability data are provided with the tests. A test-retest reliability coefficient of .96 is cited in the 5th edition of The Psychology and Teaching of Reading (Schonell & Goodacre, 1974), but no information is given concerning the numbers or the ages of the subjects from whom the data were obtained. For R4, a reliability coefficient of .92 is cited, but again there are no details of how it was obtained. Without further information, the reliability coefficients are meaningless. A test could have essentially zero reliability within one grade level (i.e/. be unable to consistently distinguish between the different levels of performance within the grade), but still have high reliability in a sample with a wide range of grade levels (because it can distinguish between, say, the performances of grade 2 and grade 8 pupils). Reliability data without some information about the nature of the sample cannot be interpreted

No data on validity are provided for these tests in any of the testing materials, or in Schonell's books. As a word recognition test, R1 has face validity, and some credence is added by the ordering of the words on the basis of their difficulty for children in the age ranges of the test. As discussed previously, if R2 measures comprehension, it is comprehension at the most literal level, since the test contains only questions requiring retrieval of facts presented in that passages. And it is extremely difficult to see what is measured by R3 and R4, which contain a heterogeneous mixture of short-answer, multiple-choice and fill-in-the-blank items, as described previously. No rationale is provided for the selection and construction of items, and it is difficult to imagine post-hoc what rationale there might have been. Some of the passages have become dated ("In some cities coloured lights are used to direct the cars at cross streets"), and the background knowledge presumed to be brought to the test is somewhat oriented toward British childre .

For all tests the only interpretive score provided is reading age, which detracts from their usefulness considerably. As we have shown elsewhere, the concept of reading age is of doubtful utility, even when the test provides a good measure of a well-defined concept, and for this reason has been discarded by most measurement experts. When what is being measured is as unclear as it is for the Schonell tests, expressing scores as reading ages will surely provide more misinformation than information.

Norms

For all four tests, tables are provided which purport to convert raw scores on the tests into reading ages. Since, as we have pointed out, the tests are very different, it is unlikely that whatever they do measure can be converted into a common metric, but this is what they claim to do. Information as to how the reading ages are derived is not provided with the tests, and is in fact extremely difficult to track down. All appear to have been obtained from the testing of children in Salford, England, which is described by Bookbinder (1970) as "largely working class." Bookbinder also provided data which indicated that Salford children performed substantially below national norms on reading tests.[2] Therefore the Salford data appear to be an unsound basis upon which to build test norms. Nevertheless this is what was done. Further details can be extracted from The Psychology and Teaching of Reading (Schonell & Goodacre, 1974, p. 216) and from Bookbinder's article. The accounts given are incomplete and conflicting, but it appears that R1 was normed on either 298 or 300 children aged up to seven and a half, in 1968, and on another 10,000 children aged from six and three quarters to eleven and three quarters, in 1971—all from Salford. The reading ages have been extrapolated upwards, since the table gives reading ages extending to 12.6+. Because Salford children scored below the national norm on intelligence tests (not cited), the reading scores have been adjusted by taking the 67th percentile of Salford children as the national norm on the reading test! In short the norms are inadequate and out-of-date, and their potentially most useful function (as local norms for Salford) has been destroyed by the tampering which could only be described as an attempt to create bargain-basement national norms. There is no reason to imagine that the norms for R2, R3 and R4 are any better, since the information provided is even more skimpy. R2 appears to be normed on a sample of 512 cases, tested (it appears) some time in the 1940's, but no details can be located. For R3 and R4 it is stated that "1865 cases were used to obtain approximate norms," but no details concerning time and place or the ages of the pupils are provided. Apparently these "approximate norms" were obtained before World War II. After the war, new norms were published which "differed materially" from the pre-war norms, but no particulars were provided.

[2] Even this information probably should be taken with a grain a salt. The national norms were obtained from a different test, and "equivalent reading ages" compared. And the Salford sample consisted of 29 boys and 21 girls.

In short, the norms provided are the least adequate which could have been chosen (reading ages), they are wildly out-of date, they are irrelevant to Australian conditions and unrepresentative of British conditions, and their documentation is shoddy, inaccurate, and incomplete. Their only virtue is that if studied carefully, they give clear implications of the psychometric qualities of the tests--for example, on R3, a change of 1 in your raw score (which is easy enough to achieve--e.g., one lucky guess) can boost your "reading age" by up to eight months. Teachers who wish to use the tests would do better to ignore the norms tables completely.

Convenience

The tests are convenient to use and score, although R1 and R2 require individual administration.

General Evaluation

We could not recommend the use of these tests. Their standardisation is woefully inadequate, and any moderately-competent teacher and construct a set of exercises which would be as good, and should be encouraged to do so.

The Standard Reading Tests

J.C. Daniels & H. Diack

Publisher: Chatto and Windus

1958

Introduction

The following twelve tests comprise the Standard Reading Tests:

Test 1 - The Standard Test of Reading Skill
Test 2 - Copying Abstract Figures
Test 3 - Copying a Sentence
Test 4 - Visual Discrimination and Orientation Test
Test 5 - Letter Recognition Test
Test 6 - Aural Descrimination Test
Test 7 - Diagnostic Word Recognition Test
Test 8 - Oral Word Recognition Test
Test 9 - Picture Word Recognition Test
Test 10 - Silent Prose Reading and Comprehension Test
Test 11 - Graded Spelling Test
Test 12 - Graded Test of Reading Experience

This battery of tests is intended to be diagnostic in nature. The tests, directions for their administration and interpretation and the rationale behind them all appear in the book The Standard Reading Tests (Daniels & Diack, 1958).

The purpose of the battery is to provide detailed information on a student's reading ability so that the teaching programme can be improved. The tests are constructed according to Daniels and Diack's beliefs that (1) there are separate skills within the broader skill of reading which when put together enable the act of reading to take place, (2) reading is "translating the letters of words, in a given order, into sounds that have meaning" and (3) the main emphasis in teaching reading should be on helping the child determine sound-symbol correspondence.

Daniels and Diack describe the two main facets of reading as being reading skill and reading experience. Although the authors do discuss these terms, it is difficult to determine what they mean by them. It seems that, for the most part, reading skill is equated with methods used by the child to provide the appropriate oral version of a written word. Reading skill, then, involves blending, syllabisation, knowledge of sound-symbol correspondences, etc. Reading experience, on the other hand, involves the child's ability to get the meaning of language based on his/her past linguistic/reading experiences. Daniels and Diack imply in their introduction that although reading skill (word recognition) and reading experience (comprehension) are related, they should be conceptualised, measured, and treated for practical purposes separately.

Eleven of the tests in the battery are tests of reading skill; Test 12 is a test of reading experience. We shall review only Tests 1 (Reading Skill) and 12 (Reading Experience) here as they seem to be the two most commonly used in the schools.

Description and Stated Purpose

Test 1: The Standard Test of Reading Skill is designed to be administered individually and is composed on 36 sentences in question form. Each sentence is reproduced on a separate page of the test book. The child's task is to read each sentence aloud and answer the question posed. (Questions 27-36 have four possible answers provided which the child must also read.) Daniels and Diack state, "The questions have been carefully chosen so that once they have been read correctly, the child is almost certain to give the right answer."[3] Children are marked only on their oral reading of he words, not on the answer they give. Thus, the SRT1 is essentially a test of word recognition in context. Results on SRT1 are meant to indicate the extent of the child's mastery of the skill (as defined above) of reading.

Test 12: The Graded Test of Reading Experience consists of 50 sentences in which the child must select, for a word in the sentence, one of four alternatives which have been provided.

Example: 11. Grass is (blue, green, white, red).

Test 12 is a group test which is meant to be administered to children who score very highly on SRT1. The authors state that Test 12 tests reading experience (as defined above) i.e., "The extent to which reading skills have been used in practice. It is as much a vocabulary test as a reading test." Thus, the purpose is still quite vague. At times it is suggested that SRT12 is a test of comprehension: "A child who scores low marks on Test 12...is meeting some difficulty in comprehending material read silently" (p. 209).

Prior to examining other aspects of the tests, it seems important to discuss in some detail the purposes of The Standard Reading Tests. Daniels and Diack have a problem with their distinction between reading skill and reading experience. They tacitly recognise the problem when they maintain that reading skill is something more than mere "word-calling", but because they conceive of reading as:

[3] In general this seems to be the case although the construction of some sentences ("Has a cup a lid?") and the concepts in some sentences (windmills) may be beyond the competence of some children.

sight------- sound---- meaning
(p. 8)

they are left with the notion that reading skill is what it takes to obtain "sound" from "sight" and that reading experience is what enables the reader to attach "meaning" to "sound". Inevitably, then, they equate reading skill with word recognition skills and reading experience with reading comprehension. Furthermore, there is always the implication that progress through the hierarchical levels of phonics leads to progress in the ultimate goal of reading, comprehension.

The authors have actually created a neat little package with their model of the reading process, the *Standard Reading Tests* which are based upon that model, and the *Royal Road Readers* which are also based on the same theoretical framework. Throughout they consistently conceive of development in reading as learning a series of hierarchically ordered sound-symbol correspondences. Ncwhere is this more evident than in their assessment of the results which would be obtained from the SRT1. Daniels and Diack cast aspersion upon the concept of reading age, saying that it provides no qualitative interpretation of a child's reading.[4] Instead, they recommend that a series of reading standards be used to interpret performance on their tests. These six standards are based on the notion of increasing ability to use phonics in reading, viz., increasing reading *skill* in Daniels and Diack's terms. The six standards are as follows:

Standard 0 - Very little functionally-operative reading ability.

Standard I - Fairly high degree of understanding that letters stand for sounds and that the order of letters in words is important.

Standard II - Know the letters functionally in words and can recognise quite complex consonantal blends. These children "need plenty of practice with properly graded readers, e.g., *Royal Road Readers*."

Standard III - Fairly high degree of understanding of the fundamentals. Now have to add to their repertoire the more complex phonics rules (digraphs, distant modification of vowels, etc.)

Standard IV - Have mastered most of the basic skills in reading.

Standards V - Have mastered all the skills of reading.

[4] Yet they provide a chart for converting raw scores on their tests to reading ages and provide reading age equivalents for their reading standards.

and VI All they require is wider reading experience.

Thus they have attempted to provide criterion-referenced interpretations in addition to the noun-referenced interpretation which are given with reading ages. The standard attained indicates to the teacher where to slot the child in on a phonically organised reading scheme (such as the *Royal Road Readers*).

The problem caused by the authors' conceptualisation of reading shows up most clearly when they discuss the interpretation of the results from the SRT 12, The Graded Test of Reading Experience. Remember that Daniels and Diack recommend that only children who score highly on SRT 1 should take Test 12. Such children would, then, according to the authors, have virtually mastered "reading skill." Yet clearly some children who score well on Test 1 score poorly on Test 12 and thus are "meeting some difficulty in comprehending material read silently." The authors go on to say, "What causes such difficulties is not easy to determine."

It follows that Daniels and Diack's instructional recommendations are necessarily weak because of the way in which they have attempted to define reading. There is always the implication that reading instruction should proceed by first teaching phonics so that children can unlock 'the sound from the sight.' According to this conception, once phonics is mastered, comprehension will follow. We shall not discuss the shortcomings of this model here; many excellent accounts of its deficiencies are readily available (e.g., Goodman & Goodman, 1977; Smith, 1982). But clearly if "reading skill" is necessary for the development of "reading experience," it is surely not sufficient.

Thus, it would seem that, despite their hesitation to do so, Daniels and Diack have essentially defined reading skill as skill in phonics and reading experience as skill in comprehending. This is how we have interpreted the authors' ideas, and we shall continue our evaluation of the SRT 1 and SRT 12 on this basis so that the reader will find it relatively consistent.

Appropriateness of Items to Stated Purpose; Item Quality

Test 1: The items are ordered according to *a priori* estimates of difficulty, and evidence is cited that this corresponds closely with the statistical order of difficulty The nature of the difficulties introduced throughout the test relates to the descriptive information contained in the *Reading Standards* but the links are not made explicit nor is any empirical justification provided in support of the *standards* interpretation. The questions are, however, straightforward enough, and do not seem to have suffered greatly with the passage of time.

Test 12: Time has been less kind to this test. Many of the items refer to concepts or experiences which are not longer familiar to children, witness the following (the intended answer is in brackets):

2. If you use a pen, you also need (ink).
6. Coal is usually (black)
14. Shoes are usually made of (leather)
16. Men's socks are usueally (knitted)
22. A steam engine usually runs on (rails)

For others, the answer would vary according to time and place:

19. Most ouses in this country today are lit by (candles, electricity?)
34. In this country, the commonest fuel used for house fire is (wood, oil, coal?)

Item 28 may have been overtaken by changes in common usage:

28. A place where talking film is shown i^ called a (theatre, cinema).

Finally, in the 50 questions, there are 11 references to male characters, and only one to a female, that being

The most important female participant in a wedding is the (groomsman, bridegroom, mother, bride).

This sort of imbalance is less acceptable today than it was in 1958. In general, it would have to be said that the items in test 12 have become dated, and then suitability for children in Australia in the 1980's is doubtful.

Passage Dependence

Not applicable to either test.

Reliability and Validity Evidence

None is provided, nor is reference given to studies which provide such evidence.

Norms

"Some teachers who use the tests may wonder why no precise norms and Achievement Ages are given for any of the tests except the Standard Spelling Test and the Reading Experience Test. The reason is that the others are tests of skills which must be completely mastered if normal progress is to be m^de" (Daniels & Diack, 1958, pg. 12). The argument is a reasonable one, and it is a pity, given the strictures against the use of reading ages (Daniels & Diack, 1958, pp. 11-12) that norms are provided in the form of reading ages and reading quotients only for Tests 1 and 12. Unfortunately no information is provided about the standardization sample, so it is impossible to judge the accuracy or applicability of these norms.

The interpretation of SRT 1 scores in terms of Reading Standards is not norm-based, and given the somewhat tenuous relationship between the item content and the interpretations provided, would require some support in the form of research evidence.

Convenience

Test 1 must be administered individually; it is easy to score, however. Test 12 is very convenient: it may be administered as a group test and is also very easy to score. The instructions on both tests are clear, and no special training is needed to administer them.

General Evaluation

The Standard Reading Tests are based on a particular notion of how children read and learn to read, a notion which in many ways fails to take account of a complex cognitive process. As a result, the SRT1 and SRT12 give the appearance of telling a teacher more than they actually can about a child's reading. The Reading Standards used to interpret scores from the SRT1 account for only a certain aspect of the operations involved in reading, and thus the pedagogical recommendations which accompany these standards are narrowly focused and could be misleading.

Furthermore, there are no data provided in the test handbook regarding the reliability or validity of the instruments, nor is there any reference to where such data may be found, if they are available at all.[5] Writing of the Standard Reading Tests in The Seventh Mental Measurements Yearbook in 1972, M. L. Kellmer Pringle said,

> As has been noted by a previous reviewer, data on reliability and validity remain conspicuous by their absence. Nor is there any indication in the manual that they can be found elsewhere. This is a serious omission, all the more so as the battery was published some twelve years ago.

It is now over twenty years since the tests were published, and still we have not a bit of information. Furthermore the extent to which the items in SRT 12 have become dated, indicated that the test itself is in serious need of revision.

[5] An extensive review of sources failed to locate any such data.

In summary, then, the SRT1 and SRT12 have little to recommend them. They are based upon a narrow theory of the reading process, have no information as to their reliability or validity, and contain simplistic instructional recommendations. We suggest you look elsewhere for an adequate standardised reading test.

References

American Psychological Association. *Standards for educational and psychological tests*. 1974.

Bookbinder, G. E. Variations in reading age norms. *Educational Research*, 1970, *12*, 99-105.

Bormuth, J. Literacy in the classroom. In W. D. Page (Ed.), *Help for the reading teacher*. Urbana, Illinois: NCRE/ERIC, 1975.

Cronbach, L. *Essentials of psychological measurement*. New York: Harper & Row, 1970.

Goodman, K. S. & Goodman, Y. M. Learning about psycholinguistic processes by analyzing oral reading. *Harvard Educational Review*, 1977, *47*, 317-333.

Haney, W. & Madaus, G. Making sense of the competency testing movement. *Harvard Educational Review*, 1978, *48*.

Pringle, M. L. K. Review of the *Standard Reading Tests*. In O. K. Buros (Ed.) *The seventh mental measurements yearbook*. Highland Park, N.J.: The Gryphon Press, 1972.

Royal Road Readers.

Schonell, F. *Backwardness in the basic subjects*. Edinburgh: Oliver & Boyd, 1948.

Schonell, F., & Goodacre, E. *The psychology and teaching of reading*. (5th Ed.) Edinburgh: Oliver & Boyd, 1974.

Smith, F. *Understanding reading*. (3rd ed.) New York: Holt, Rinehart and Winston, 1982.

Further Reading

Additional information on standardised testing and the teaching of reading can be obtained from the following sources:

Buros, O.K. (Ed.) *Reading tests and reviews, I and II*. Edison, New Jersey: Gryphon Press, 1968, 1975.

Farr, R. *Reading: What can be measured?* Newark, Delaware: International Reading Association, 1970.

Fraser, R., & Fehring, H. Testing 1, 2, 3! In Education Department of Victoria, Curriculum and Research Branch. *Reading is for meaning*. Melbourne: Education Department of Victoria, 1979.

Fehring, H. *Let's look at standardised reading tests*. Occasional Paper No. 12, Curriculum and Research Branch. Melbourne: Education Department of Victoria, 1979.

Hogben, D. Grouping primary school pupils for instruction in reading on the basis of scores from standardised reading tests. *The Australian Journal of Education*, 1978, *22*, 295-302.

Macginitie, W.H. (Ed.) *Assessment problems in reading*. Newark, Delaware: International Reading Association, 1973.

Pumfrey, P.D. *Measuring reading abilities: Concepts, sources and applications*. London: Hodder and Stoughton, 1977.

Rowley, G. Reading age and associated myths. *Australian Journal of Reading*, 1980, *3*, 76-85.

Strang, R. *Diagnostic teaching of reading*. (2nd ed.) New York: McGraw-Hill, 1969.

Schreiner, R. (ed.) *Reading tests and teachers: A practical guide*. Newark, Delaware: International Reading Association, 1979.