DOCUMENT RESUME

ED 116 149 CS 002 313

AUTHOR Braun, Carl
TITLE The Standardized Test: Uses and Abuses.
PUB DATE 75
NOTE 20p. Paper presented at the Annual Meeting of the Transmountain Regional Conference of the International Reading Association (2nd, Calgary, Alberta, November 13-15, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Achievement Tests; Behavioral Objectives; Cognitive Processes; Cognitive Tests; *Criterion Referenced Tests; Intelligence Tests; Listening Skills; Measurement Techniques; *Norm Referenced Tests; Reading Skills; *Standardized Tests

ABSTRACT

Standardized norm-referenced tests have been much maligned in recent years. They differ from criterion referenced tests in that the latter involve assessment in comparison to an absolute standard or specific performance objective while in a norm-referenced test assessment is made in comparison to other students taking the same test. Standardized tests can be abused in some of the following ways: failing to recognize limitations within the testing situation that obscure the "true" level of competence; making decisions on the assumption that the score derived from the test tells all; making decisions on the assumption that reading and intelligence tests measure exclusive domains; assuming that a standardized test can give specific direction to an instructional program; interpreting results without reference to the composition of the norm group; making the assumption that anyone below the 50th percentile is a disabled learner; and treating a grade score on a reading test as a functional reading level. However, standardized tests do have many positive uses: as an accountability check, as a screening device to determine further diagnostic needs, and as a way to generate hypotheses regarding instructional needs. (MKM)

The Standardized Test: Uses and Abuses

Carl Braun

The University of Calgary

Paper presented at the Second Transmountain Regional Conference (International Reading Association)

November 15, 1975

THE STANDARDIZED TEST : USES AND ABUSES

The issue surrounding the use and abuse of standardized tests has frequently proven more an emotional than an intellectual one. As so often happens with emotional issues, standardized test discussions tend to dichotomize people. This is reflected in debates like "criterion vs norm-referenced testing", "to brain watch or to educate", "quantitative vs qualitative aspects of education". The ultimate in emotional outbursts is exemplified in the title of a kit recently developed by the National Council of Teachers of English, "A First-Aid Kit for the Test-Wounded". A close rival is reference to tests as "prejudicial educational traps" in an article entitled "Shot Down by the Tests" (Skinner, 1968, p. 13). The defiant tones of one camp denouncing the possibility that a worthwhile educational objective can be quantified is matched only by the hushed tones of the other camp who view the score derived from a succession of squiggles comprising nothing less than a magic number - a magic number that can be viewed in absolute terms. Risking the role of a "fence-sitter", I submit that tests, with a few notable exceptions, are neither good nor bad. Like atomic energy, standardized tests can serve useful functions; they can also easily be used against a child. This point must be emphasized - a test will never "wound" anyone nor will it "trap" anyone; the decision of "wounding" or "trapping" lies within the prerogative of the test user.

In this discussion I propose to summarize very briefly the differentiation between norm- and criterion-referenced testing; abuses of norm-referenced (standardized) tests and their uses. The reversal of the key words "abuses" and "uses" in no way represents a bias, but rather a concern to end the discussion on a positive note.

DIFFERENTIATION BETWEEN CRITERION- AND NORM-REFERENCED TESTING

Tests are designed to answer questions that educators raise. "What is Jason's specific problem in reading?" "How well is my class doing in relation to classes elsewhere?" "Did I do as good a job teaching this year as last year?" "Should I shift emphasis in my teaching program in word identification skills?" Which of the questions are to be answered will be determined by the choice of type of test- criterion or norm-referenced.

Space does not permit a detailed discourse on criterion-referenced testing nor is it necessarily within the interest of the present discussion to do so. However, a brief differentiation between the two types of testing is presented.

The major difference between norm- and criterion-referenced tests lies in the way the items are developed and selected and the way in which the test is to be interpreted. Criterion-referenced tests imply that a student is assessed in comparison to an absolute standard rather than in comparison to other students taking the same test (Good, Biddle and Brophy, 1975, p. 155). These tests are developed to yield measurements that can be interpreted directly in terms of specific performance objectives. For example, "The learner can identify the main idea of a paragraph 95 percent of the time". Note that there is no implication as to whether this is good or bad. Only the teacher can decide this. Note also that the conclusion based on testing the objective clearly implies direct instructional needs. From this standpoint the criterion-referenced test is useful in aiding 'on the spot' decisions regarding instruction.

Since criterion-referenced tests have their base in performance of objectives, the tests are subject to the same weaknesses and strengths as

performance objectives themselves. Such testing involves breaking down a subject area into small instructional units so that all students can master a commonly agreed upon set of skills. One of the major obstacles facing proponents of criterion-referenced testing has been the question of agreement on the domain of skills to be included. Another problem has been the lack of agreement on criterion levels. (Is 80 percent or 100 percent mastery minimum performance level?) Perhaps the greatest hazard has been the temptation to test those skills which submit readily to statements in performance terms.

Norm-referenced testing is not concerned with 80 or 100 percent mastery levels; it provides meaning to a student's score only by comparison of his test performance with that of others on the same test rather than comparison against an absolute standard. Whereas, criterion-referenced testing denotes high scores, or "bunching" scores at the top level, norm-referenced testing spreads students' scores as far as possible. This is accomplished by posing questions that roughly 50 percent of the students can respond to correctly and that are responded to correctly more often by students who attain high total scores than by those who achieve a relatively low total score. Norm-referencing, by definition, denotes that equal numbers of students in the norm sample score above and below grade levels.

Compared to criterion-referenced tests, norm-referenced tests are typically, although not exclusively, designed to evaluate more global aspects of the curriculum and thus have less relevance to immediate instructional application.

In summary, then, what is it that we are attempting to accomplish in standardized (norm-referenced) testing? Basically, we are observing a sampling of a student's behavior from which we are making an estimate of his "true" level of competence. This estimate may be used for predictive purposes,

and to determine what changes in curriculum and instruction should be made. Whatever the purpose, the assumption generally is that testing conditions, the test, and the interpretation of the test are optimum. Blind acceptance of scores from standardized tests has often led to varying degrees of abuse.

ABUSES OF STANDARDIZED TESTS

Failure to recognize limitations within the testing situation that obscure the "true" level of competence.

Any condition within the test situation that reduces the reliability of a given test masks or obscures the competency level attained by the testee. Perhaps the most critical factor affecting the reliability has to do with anxieties generated within the testee. Seiler (1970) attributed fear of test situations as one of the anxiety-producing variables. The aura or mystique shrouding the test, he felt, created anxieties which accounted for reduced productivity. He reported one survey which revealed that some adult applicants for testing thought that taking a test was like a medical examination requiring various stages of undress. In the same survey one applicant interpreted "test battery" as demanding knowledge of electricity.

Anxieties spring from the most unsuspecting sources. A precocious kindergarten child overheard the psychologist tell the teacher that he would be back in the afternoon to "wind up" the rest of the testing. The parent of the child called the school during lunch hour reporting her child's reluctance to go to school because a stranger was going to "wind her up".

A further factor that affects both reliability and validity of tests is the format in which the item is cast. Comprehension, for example, is measured in many ways. If a child is working the items on the Stanford Reading

Test, he will be given a cloze paragraph with response choices at the bottom of the paragraph. If he is doing the comprehension test from the Monroe-Sherman Diagnostic Reading Test, he will be confronted with a question first, will then read the paragraph, and then circle one of several choices to answer the question. If he is tested on comprehension with the Durrell Analysis of Reading Difficulty he will simply try to recall the ideas he has read in a paragraph.

Misinterpretations of test results frequently stem from another type of item which requires the testee to circle one of four or five words that best corresponds with a given pictorial stimulus. Frequently, the obscurity of the picture or the experiential background of the child limits his ability to identify the picture correctly and, of course, ultimately his ability to circle the correct word.

Frequently tests are interpreted without reference to the response level required by the testee. For example, there is a substantial difference between mere recognition and identification. If the testee is required to recognize the word "funny" in the series, "fair", "funny", "flew", "folly" his response level is different from the requirement that he identify the word "funny" without aid.

Closely linked with format of the test item is the inclusion of value or direct experience-based items. Schiller (1974) refers to an item on the WISC which states, "If your mother sends you to the store for a loaf of bread and there is none, what do you do?" The child who answers, "I go back home", is considered to be intellectually inferior to the child who says, "I go to another store". The point is that many children in rural areas have only one store to go to. It is also conceivable that a child in a city gets instructions to go to a specific store and feels that to go to another could be interpreted as disobedience. Dreskin (1965) gives an example of how choice of vocabulary

on intelligence tests often favors children from "better class homes". When children from "better-class" homes and "lower-class" homes were confronted with the analogy, "Symphony is to composer as book is to (paper, sculptor, author, musician, man)," the first group scored correctly 81 percent of the time against 52 percent for the second group. When the analogy was re-worded to "Baker goes with bread as carpenter goes with (saw, house, spoon, nail, man)," the two groups scored evenly in checking "house". If the objective of the item is to ascertain the learner's facility to deal with analogy, the word "symphony" certainly appears to have set up a barrier to differentiate those learners who can handle analogy from those who can not.

A further source of misinterpretation can arise from the fact that some may not be familiar with formal test-taking behavior. Ruddell (1974) attributes low achievement scores of some children to:

> pupil unfamiliarity with labels and concepts used in test situations, i.e., failure to understand the task required to respond in test items; and unfamiliarity with labels and concepts being evaluated by instrument (p. 384).

A final source of misinterpretation may arise from failure to recognize that certain items on comprehension tests can be answered without dependence on the written passage. Tuinman (1973) found that probabilities of correct responses on test passages not read by students in grades four to six were well above the expected chance level. Average probabilities of correct responses with no passage present ranged between .32 and .50.

In summary, failure to recognize limitations within the testing situation can well obscure the testee's "true" level of competence. The varying formats, content and test conditions can only too easily lead to the situation described by Dreskin (1965). He reports the IQ scores of a girl whose father was in the armed forces and had been stationed across Canada. The parents were understandably confused by the fact that their daughter's

intelligence ranged all the way from low average to superior depending on where she had been tested : 110 in British Columbia, 90 in Manitoba, 115 in Ontario and 125 in New Brunswick.

Making decisions on the assumption that the score derived from the test tells all.

A test score can be interpreted only in the light of the degree to which the items sample the domain of the construct represented. For example, a silent reading test does not tell us nearly everything about the testee's reading. Reading is a highly complex cognitive and affective process. What we are getting from the student's silent reading is a small sampling of the product of his reading - very little about the process. Again, item format has some bearing on this. It would appear that analysis of a test cast in cloze format yields more information on process than questions answered subsequent to paragraph reading. The anomalous nature of part scores on reading tests is nowhere more evident than it is in the case of reading comprehension tests (Traxler, 1970. p. 223). Many comprehension tests consist largely of factual questions while others emphasize aspects of critical, inferential and creative reading. Meaning can be attributed to the learner's score only in the light of a close examination of the test domain.

Even rate of reading is not the simple procedure it appears to be superficially (Traxler, 1970. p. 222). We have to be concerned with rates rather than one rate. Content is an important determining variable of rate from the standpoint both of concept familiarity and load and personal interest or motivation. Further, interval of time is an important factor in determining reading rate. Traxler recommends a sampling of at least three to five minutes.

Making decisions on the assumption that reading and intelligence tests measure exclusive domains.

It is not uncommon to hear school personnel reflect on the cumulative report of a child in the following manner:

> I don't understand; Charles had an IQ of 110 when he was tested in grade two, 101 in grade four, and now two years later he is down to 91, almost a candidate for a special class. No wonder he has trouble reading.

There are at least two related problems. First, if the child has been given a group test which is likely, his inability to read is going to reflect cumulatively in the intelligence tests. This does not take into account additional problems of failure complexes and increasing lack of motivation.

Further, intelligence tests sample content closely akin to reading comprehension tests. After all, reading is thinking. Traxler feels that the better and more searching the reading test is, the greater this limitation becomes (p. 224). So, scores on reading tests really represent a composite of reading and intelligence.

Assuming that a standardized test can give specific direction to an instructional program.

Because of the highly generalized nature of most standardized achievement tests, they do not measure the specific objectives for a particular student (Ruddell, 1974. p. 384). At best this very global assessment can give very general directions for instruction. Two possible exceptions come to mind. First, if, say, a reading comprehension test samples a wide spectrum of

comprehension tasks ranging all the way from factual recall to inferential reading, a careful item analysis will aid in revealing specific instructional needs. Second, if the test is designed to yield a diagnostic profile, specific instructional trends can be revealed. Generally, however, group tests are designed to measure the achievement of groups rather than the educational placement of individuals.

Interpreting results without reference to the composition of the norm group.

It has been mentioned earlier that an individual's score on a standardized test is interpreted in comparison with the performance of the norm group. Ruddell (1974) states explicitly that:

> Because the "objective" scores students receive on a standardized reading achievement test are determined by the norm group to whom they are compared, these tests tell little about student achievement unless this norm group is completely and accurately defined. Boards of education, the community, parents, and even professional educators often misinterpret achievement test scores for this very reason (p. 385).

Making the assumption that anyone below the 50th percentile is a disabled learner.

As mentioned earlier, the very nature of a norm-referenced test means that scores will be spread out or to put it another way, that the test will differentiate between weak and strong students. If we administer a reading achievement test to a group of students similar in composition to that of the norm group, we can expect approximately half of the students to fall below the 50th percentile.

Giving special instruction to enhance performance on the test.

Pressure generated through lack ~~of~~ understanding has frequently resulted in specific instruction to raise scores on achievement tests. This form of corruption negates any value to be gained from the test results. Again, this kind of pressure can result from a misinterpretation of the basic notion of norm-referenced tests. If, for example, superintendent X finds that School A's achievement is considerably beyond that of School B, and communicates his concern, pressure may be felt to note and select for special emphasis particular areas from achievement tests. There may be good reasons why the achievement of one school is different from the other - socioeconomic level, teacher turn over to mention only two. There is no suggestion here to discourage close examination of test results on a comparative basis to raise hypotheses about curriculum and instructional practices.

Treating a grade score on a reading test as a functional reading level.

Standardized tests suffer wide abuse from over-interpretation. Assuming that a grade score of 5.0 on a group test indicates either an independent or instructional reading level is too typical. Again, the global nature, item selection and group nature of most standardized reading tests denies such interpretation. Functional levels are best ascertained by allowing the reader "try-outs" with actual content material or by administering an informal reading inventory.

Using test results to separate students for status purposes (assignment of awards, grades, etc.)

Little needs to be said about the folly of using standardized test results for assignment of grades or promotion purposes. Again, the nature of content selection and the fact that most tests are designed to assess achievement of groups places the validity for such purposes into serious question.

Using test results for permanent grouping or streaming.

The widespread use of standardized tests for the purpose of grouping continues inspite of clear evidence of the invalidity of such practice. Perhaps, the most convincing evidence that reading achievement test scores do not differentiate specifically enough to ensure homogeneity is that provided by Balow (1962). He found in his investigation that when four classes of fifth graders were "streaming" on the basis of reading test scores, the groups still were essentially heterogeneous. When specific subskills were evaluated there was considerable overlap between the highest and lowest streams. What, in fact, happens is that the global comprehension score masks specific individual instructional needs. It is not within the interest of this paper to debate the pros of homogeneous versus heterogeneous assignment of groups. There is, however, sufficient evidence to support heterogeneous assignment even if we could determine methods of "homogenizing" groups. There is no implication here that short-term groups should not be formed for specific skill instruction. The point

is that a survey achievement test cannot adequately accomplish such a differentiation.

After such an extended discourse on the abuses of standardized testing, it seems incredible that any value could be attributed to the use of tests. This is not the case. Used for their intended purposes, considerable value can be derived from their use.

USE OF STANDARDIZED TESTS

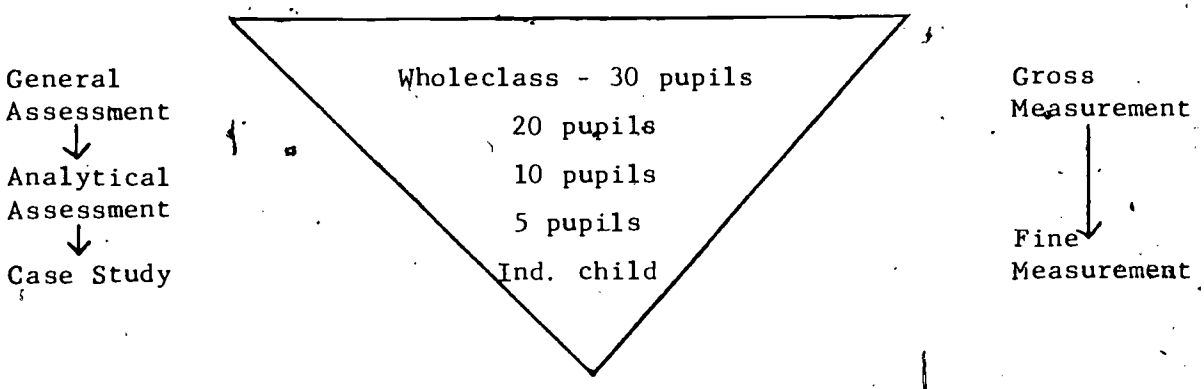
Using the test as an accountability check.

Traxler (1970) considers the most important value of a reading test, or any other standardized test, to be the definiteness that it lends to our thinking about a pupil or group (p. 226). To use a test as an external monitor adds a degree of objectivity lacking in class or even schoolwide tests. One a class, school, or system-wide basic, a standardized program can serve to provide a basis for evaluation of global aspects of the program. This applies both to ascertaining changes in achievement over a number of years as well as to determining effects of a program on a shorter term basis. The important caution is that the basis is not as solid and dependable as the "bald, bold figures" suggest (Traxler, p. 226) because of the limitations reviewed earlier.

Using the test as a screening device to determine further diagnostic needs.

A systematic approach to identification of learners who need special instruction is requisite to an efficient program. Ideally, this identi-

fication begins with gross measures applied on a whole-class basis. At this level the standardized survey test is the reasonable choice. Subsequent measurements become succesively more specific and precise for, say, a small group of children who have been screened by means of a survey test to indicate problems. Harris (1970) presents a pyramid model of successive levels of screening. The model is illustrated here:



Successive Levels of Screening (Harris, 1970, p. 95)

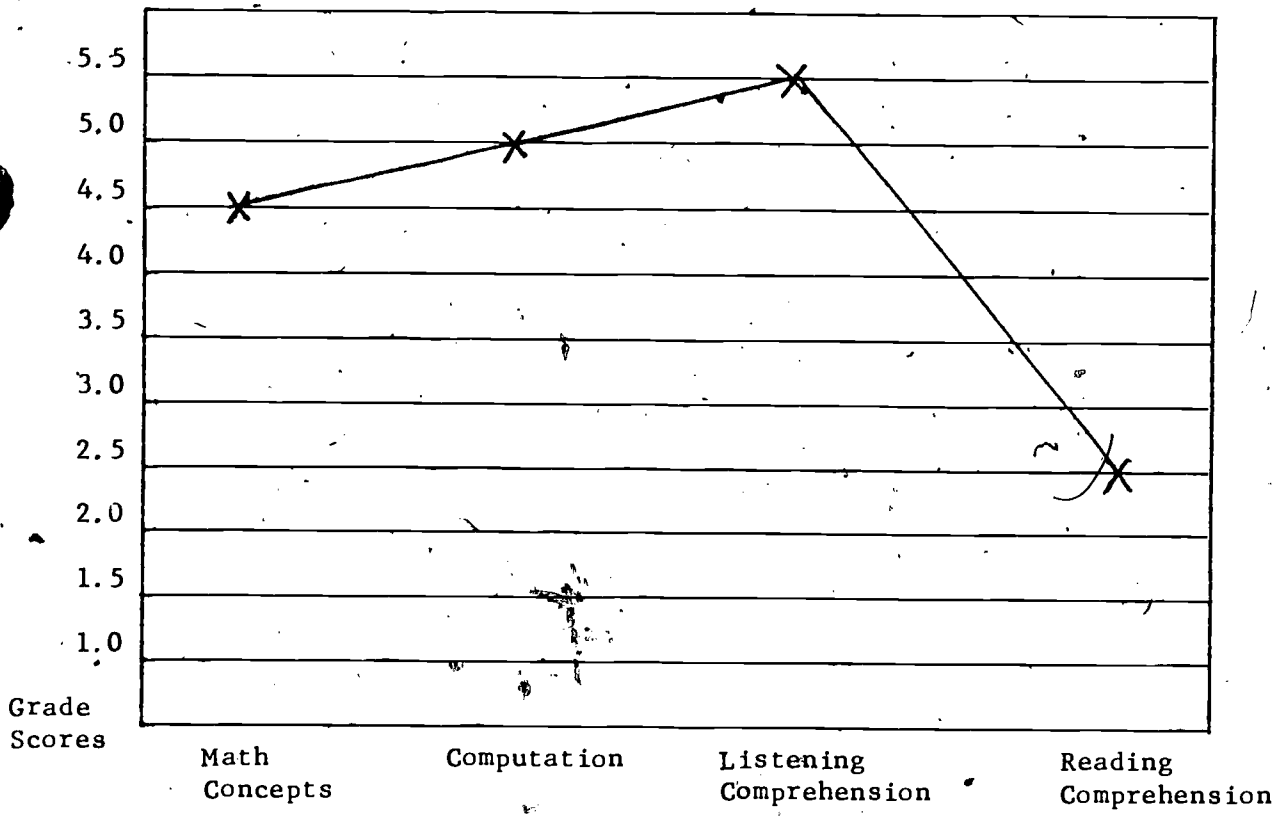
At the analytical assessment level, it is likely that a group diagnostic test would be administered. This would be supplemented with various informal diagnostic techniques. At the case study level, individual tests and observational procedures would be employed.

Using the test to generate hypotheses regarding instructional needs.

An atmosphere within a school that stimulates use of tests to generate hypotheses about instructional needs is the next best thing to an in-school research program. These hypotheses may relate to "across curriculum" or "within curriculum" concerns.

Across the curriculum; for example, the concern may be to collect data on relative achievement strengths and weaknesses in reading, listen-

ing and mathematics. Suppose a group or an individual within the class exhibits the following profile:



One thing becomes clear immediately. The learner certainly has the capability to improve in reading judging from his performance in mathematics and listening comprehension. The questions which arise might be the following: "If X is able to comprehend so well at the listening level, obviously able to process information, is it word identification skills which account for his problem in reading?" "If so, which specific skills are lacking?"

It is not uncommon to use an across-curriculum examination as a basis for determining reading expectancy levels. Otto, Smith and McMenemy (1973) suggests use of both mental age (based on group intelligence tests) and mathematics age as a basis for comparison with reading age. The follow-

ing excerpt shows how the classroom profile is set up and the information gained from it:

TABLE I:

** Comparisons of Chronological, Mental, Reading, and Arithmetic Ages of Seventh-Grade Pupils

Pupil	Chronological Age (CA)*	Mental Age (Ma)*	Reading Age (RA)*	Arithmetic Age (AA)*	Difference Between MA and AA	Difference Between MA and RA
1	12.0	14.3	15.0	13.2	-1.1	+0.9
2	12.5	16.1	15.0	14.8	-1.5	-1.1
3	12.4	15.2	15.0	14.1	-1.1	+0.2
4	11.9	16.7	14.6	14.5	-2.2	-2.1
5	12.1	11.4	12.4	12.1	+0.9	+1.0
6	11.8	12.0	9.1	11.2	-0.10	-2.11

* All ages are in years and months.
** Taken from Otto Smith and McMenemy (1973, p. 106).

Naturally small differences between M.A. and R.A. or A.A. and R.A. have to be ignored as chance level differences resulting from measurement error. On the other hand, differences in the magnitude of plus or minus 1.0 at the elementary level should be cause for careful further analysis.

Perhaps even more important is the use of tests to examine problems within a curriculum area. This is a particularly fruitful area for stimulating instructional changes. To illustrate how such changes can come about, the writer was engaged on a consultative basis in a northern reserve school. Two grade two classes had been randomly assigned to their classes. In May an achievement test was administered to both classes. The

test consisted of a Word Recognition Test (words in isolation), Comprehending Significant Ideas, and Comprehending Specific Instructions. Profiles were constructed for each individual child. An analysis revealed that in class A, 17 out of 24 children were as high or higher in word recognition as in either of the comprehension tests. In class B the reverse was observed - 19 out of 26 pupils were at least as high in comprehension as in word recognition. An examination of these results led to serious discussion about the instructional program carried out in classroom A. The changes in instruction the following year were most apparent. The results at the end of the year confirmed the impact of the instruction.

A standardized reading survey test can reveal considerable information if it includes a wide range of comprehension questions. If these questions are then clustered (e.g., #'s 3, 7, 11, 15, etc. are main idea; #'s 2, 4, 8, 14, 19, etc. are implied meanings and so on), the teacher can determine individual instructional level but can also get an indication of areas where her instruction tends to leave gaps.

Zehm (1975) reports findings of research which resulted from the discovery that second grade students in San Francisco schools dropped well below the national norms in reading. The investigation isolated four schools where the reverse was true - reading scores were above the national average - to determine the sources of success. Class size was not the variable; nor was the number of minority students. In fact, the researchers found 40 to 100 percent minority students in these classes. Further investigation revealed that neither technique, capital outlay or method was the key. Methods, in fact, varied from highly structured approaches to the more flexible style of the open classroom (p. 25). The key to success was found in the attitude of the teachers. They were enthusiastic, positive, optimistic about their students' potential, and emphasized reading in every subject.

There is no need to summarize in an attempt to debate whether abuses of tests outweigh their uses. This should, in fact, never become an issue. We know the value that tests can have; we only need to avoid the widespread abuse of these instruments.

References

1. Balow, Irving. "Does Homogeneous Grouping Give Homogeneous Groups?" Elementary School Journal, October, 1962. pp. 28-32.

2. Dreskin, N. "What I.Q. Tests Don't Tell Us". The Winnipeg Tribune Weekend Magazine, February 13, 1965.

3. Good, Thomas L., J. Biddle, and Jere E. Brophy, Teachers Make A Difference, Holt, Rinehart and Winston, 1975.

4. Harris, Larry A. "Evaluating a Reading Program at the Elementary Grade Level," in Roger Farr (ed.), Measurement and Evaluation of Reading, Harcourt, Brace and World, 1970.

5. Otto, Wayne, Richard A. McMenemy and Richard J. Smith. Corrective and Remedial Teaching, Houghton Mifflin Company, 1973. p. 106.

6. Ruddell, Robert B. "Achievement Test Evaluation - Limitations and Values," in Robert B. Ruddell (ed.), Resources in Reading-Language Instruction, Prentice-Hall, Inc., 1974. pp. 383-386.

7. Schiller, Janine. "The Abuse of Standardized Testing," California English, September, 1974. pp. 10-11.

8. Seiler, Joseph. "Preparing the Disadvantaged for Tests," Manpower, Vol. 2, No. 7, July, 1970. pp. 24-26.

9. Skinner, Vincent P. "Shot Down by the Tests," Maine Teacher, December, 1968. p. 13.

10. Traxler, Arthur E. "Values and Limitations of Standardized Reading Tests," in Roger Farr (ed.), Measurement and Evaluation of Reading, Harcourt, Brace and World, Inc., 1970. pp. 220-229.

11. Tuinman, Jaap J. "A Large Scale Study of the Passage Dependency of Items in Five Comprehension Tests," Paper presented at the International Reading Association Convention, Denver, 1973.

12. Zehm, Stanley J. "Teacher Expectations: A Key to Reading Success," Reading Improvement, Vol. 12, No. 1, Spring, 1975. pp. 23-26.