DOCUMENT RESUME

ED 117 197 95 TM 005 057

AUTHOR Joselyn, E. Gary
TITLE An Introduction to Standardized Testing for Teachers and Administrators.
INSTITUTION Educational Records Bureau, Framingham, Mass.; ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-55
PUB DATE Sep 75
CONTRACT NIE-C-400-75-0015
NOTE 12p.
AVAILABLE FROM Educational Records Bureau, Framingham, Mass. ($2.00)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Administrative Personnel; Aptitude Tests; Elementary Secondary Education; *Guides; Norms; Scores; *Standardized Tests; Student Evaluation; *Teachers; Test Interpretation; Test Validity

ABSTRACT

Most problems with tests have to do with their use, misuse, or lack of use. Test scores can be of value to teachers who know how and how not to use them. The purpose of this booklet is to provide a brief overview of standardized testing and to explain some of the commonly used terminology. Topics discussed include: teacher-made and standardized tests, uses of standardized test results, types of standardized tests, test validity, scores and norms, derived scores (percentile ranks, grade equivalent, stanines), working with student profiles, and aptitude test scores. The bibliography is confined to a few readable sources which emphasize the understanding and use of tests in more detail. (RC)

AN INTRODUCTION TO STANDARDIZED TESTING FOR teachers & administrators

E. GARY JOSELYN UNIVERSITY OF MINNESOTA

This publication was commissioned by the ERIC Clearinghouse on Tests, Measurement, and Evaluation. The Clearinghouse operates under contract with the National Institute of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not therefore represent official National Institute of Education position or policy.

AN INTRODUCTION TO STANDARDIZED TESTING FOR TEACHERS AND ADMINISTRATORS

E. GARY JOSELYN
UNIVERSITY OF MINNESOTA

Testing, like so many things these days, has its extremists. On the one hand, there are those who believe all tests are worthless, unfair, and even damaging to schools and students. At the other end of the continuum are those who place blind, unquestioning faith in test scores, attributing almost magical qualities to them. The truth, I believe, lies somewhere between. Most problems with tests have to do with their use, misuse, or lack of use. Test scores can be of value to teachers who know how and how not to use them.

The purpose of this booklet is to provide a brief overview of standardized testing and to acquaint you with some of the commonly used terms. The bibliography is confined to a few readable sources which emphasize the understanding and use of tests in more detail.

Teacher-made and Standardized Tests

Teacher-made and standardized tests complement each other. Both are necessary for adequate evaluation of individual pupils and groups of pupils. Teacher-made tests are given quite often to monitor pupil and class learnings in rather specific areas that are the subject of recent classroom instruction. Their content is specific to the content of a particular classroom and reflects the specific objectives of the teacher. Standardized tests, usually given only once a year or even less often, offer comparisons with external groups, in broader achievement areas. They provide *standardized* measures and are administered under carefully prescribed conditions.

Uses of Standardized Test Results

Test results may be used for administrative, guidance, or instructional purposes.

Schools use standardized test scores for *administrative* purposes such as getting an overall picture of the level and range of abilities and achievement of the student body, placing students in special groups, and evaluating curriculum.

Guidance uses have as their principal objective greater self-understanding on the part of individual students. Test scores, often used in

one-to-one interviews with guidance counselors, help students identify their own strengths and weaknesses and make educational vocational plans.

Instructional uses are by classroom teachers for the purpose of improving and individualizing instruction.

Only instructional uses are addressed in this booklet, but teachers should be aware that standardized test scores have many uses for many audiences in addition to classroom applications.

Types of Standardized Tests

Almost all tests may be categorized as one of four kinds: aptitude, achievement, interest, or personality.

Aptitude tests are designed to measure a person's potential—that is, to predict performance at some future time, to measure what a person can learn.

Achievement tests indicate a person's present proficiency or what he *has* learned.

Interest tests are designed to help students understand their own interests and how these may relate to various occupations or courses of study.

Personality tests include a broad range of instruments that attempt to describe how persons adjust to their environment. Since classroom teachers seldom see or use interest inventories or personality tests, they are not discussed here. The primary emphasis in this discussion is on the use and interpretation of achievement test results with some brief attention given to aptitude tests.

Test Validity

The validity of a test is the degree to which that test measures what it purports to measure. Achievement tests attempt to describe what a person has learned. The validity of an achievement test, therefore, is determined by carefully examining the content of the test and making a judgment as to how adequately it samples the subject area.

Aptitude tests attempt to predict a person's performance at some future time. Thus, the validity of aptitude tests is determined by studies that investigate how closely performance on the test is related to later performance in the situation the test purports to predict.

Teachers can usually assume that those who selected a particular test for a school's testing program studied the test carefully and are

satisfied that it has good validity for use in the school. When using achievement tests, however, it is important that teachers examine the content of the test by looking at the item outline and the items themselves and make their own judgments as to how closely the test reflects the instructional goals and objectives of their subject areas.

Scores and Norms

A student's performance on a test is described by a test score. A *raw score* is simply the number of test items a student answered correctly, or this number adjusted to correct for guessing. Raw scores have little meaning in themselves because tests vary in the number of items they have and in the difficulty of their items. To give them meaning, raw scores are converted to another type of score. Any test score other than a raw score is called a *derived* score, and there are many different kinds. (Some are discussed in the following section.)

Derived scores give meaning to a student's test performance by comparing it with the performance of some known group. The known group to which the test has been given and which supplies us with a reference for evaluating the score of the individual is known as the *norm group*. Knowledge of the norm group is obviously very important for the proper interpretation of test scores. Although precise knowledge about the make-up of a norm group is of vital concern to persons charged with the responsibility of selecting a particular test for a school's testing program, normally it is not of much concern to the classroom teacher. The tests used in a school's testing program are usually chosen by persons who study the tests and the norm group thoroughly, and classroom teachers can usually trust that the norms are adequate and appropriate.

When interpreting scores, however, it is critical to keep in mind the norm group to which the scores refer:

National norms compare students' performances with those of a large group of students selected to be representative of students at the same grade level throughout the nation.

Local norms compare students' performances with those of their classmates of the same grade level in the same school system.

Sometimes scores are based on other norm groups. In addition to commonly used national and local norms, you may run across norms based upon students of a particular region or state, students of different levels of ability, or students in particular kinds of schools (large-small, urban-suburban, public-private).

Remember that *all* scores (except raw scores) are tied to some norm group and therefore describe relative, not absolute, performance.

Some Derived Scores

Three of the most commonly used derived scores are grade equivalents, percentile ranks, and stanines, which are described below.

Grade-equivalent (GE) scores: Grade-equivalent scores, sometimes called grade-level scores, represent the most common method of reporting performance on achievement tests. GE scores show the average score for students at a particular grade level. If, for example, the average raw score for all students in the norm group taking the test at the beginning of the sixth grade is 51, then 6.0 becomes the GE score for a score of 51. The first digit in the GE score is the grade level and the second is the month. A grade equivalent of 4.3, for example, represents the average performance of students in the third month of the fourth grade.

One reason for the popularity of the GE scores is that they seem easy to interpret. However, *they are often misinterpreted,* and teachers should remember the following points:

1. Different achievement batteries are written by different authors and are published by different publishers who sample different groups of students to make up their norm groups. One should not expect, therefore, that a student taking two different tests of the same kind (reading, for example) will necessarily receive the same GE score on both.

2. One should not interpret a GE score to mean that a student should be given learning materials designed for that particular level. It cannot be assumed that a fifth grade student who receives a GE score of 7.0 should be promoted to the seventh grade.

3. A common and easy-to-make misinterpretation of GE scores is to assume that identical GE scores on two different subtests represent equivalent performance on each as compared with other students in the same grade. For example, a fifth grade student who achieves a GE of 7.0 on both the reading and arithmetic tests of an achievement battery would seem to have performed equally well in both subject areas. Actually, while he got as many items right as the average student at the beginning of the seventh grade on both tests, his performance on the arithmetic test is considerably better than on the reading test *as compared with fifth grade students.* This is because the spread of scores is different for almost every subtest on an achievement battery. While it is not necessary for a teacher to know the exact amount of the difference in the spread of scores for each subtest, he should know that these differences exist and he should avoid the conclusion that equal GE scores indicate equivalent performance in two different subject areas.

4. Finally, teachers must resist the temptation to use GE scores as standards of performance. We often hear people say "Forty percent of our students are reading below grade level" or "bring everyone up to grade level." Such statements imply that it is bad if anyone scores below grade level. They reveal ignorance of the fact that GE scores represent the *average* performance of students at a particular grade level and, by definition, half of any group must be "below average." With the accountability movement gaining momentum, it is important that teachers help both parents and their fellow teachers understand that GE scores represent average performance and, therefore, cannot be used as evaluative standards of performance.

Percentile Ranks (PR): Of all the different derived scores, the PR is probably least subject to misinterpretation. A student's percentile rank represents the percentage of students in the norm group who received the same or a lower score. A PR of 65, for example, indicates that the student performed as well or better than 65 percent of the norm group. And, of course, 35 percent scored higher. It is important to keep in mind that percentile rank scores represent percentage of students in the norm group, not percentage of items answered correctly.

One problem with percentile rank scores is that they do not reflect the fact that academic achievement scores tend to bunch near the average score in the middle and spread out toward the high and low extremes. There may be a tendency, therefore, to place too much importance on percentile rank differences near the middle of the range and to place too little importance on differences near the extremes. Percentile ranks of 50 and 55 probably represent insignificant differences in performance in terms of the number of items answered correctly, while percentile ranks of 90 and 95 do represent significantly different levels of performance.

Stanines: Many achievement test reports include another type of score called a stanine. The name comes from standard scores of nine units. Stanine scores have several advantages. Each stanine value represents approximately equal ranges of scaled scores, which avoids the problem of overemphasizing small, insignificant differences in the middle of the range that could appear as large differences when expressed as percentile ranks. The statistical characteristics of stanine scores are such that one may, with a fair amount of confidence, interpret a difference of two stanine units between the scores on two tests as representing true differences in performance.

Working with Student Profiles

Achievement battery scores for individual students or groups of students are usually shown graphically on profiles which provide a visual display of a person's or a group's overall level of achievement and

particular strengths and weaknesses. Profiles of percentile rank scores may be plotted on scales on which the distance between percentile rank points is collapsed in the middle of the scale and expanded near the extremes. This helps to avoid the problem of misinterpretation of score differences in the middle and near the extremes that was discussed previously.

When interpreting a student's achievement battery profile, look first at the overall level of the scores. Although almost every student scores better in some areas than in others, there is a tendency for the scores of individuals to fall fairly close together. How does the overall level of measured achievement fit with your expectations, based on your knowledge of your students' performance in classes and on other measures?

Teachers usually find that their predictions of students' test performances are fairly accurate. But occasionally a teacher finds that scores on the profile are quite different from what was expected. It is at these times that standardized test results may serve their most useful purpose. Testing may be worth the effort and expense if even one quiet, low-achieving student shows up much higher than expected on the test, is thereby brought to the attention of the teacher, and is motivated to achieve his full potential.

Next, look at the peaks and valleys in the profile. Notice the areas in which the student seems to be particularly strong or weak and consider their implications for planning and instruction so the strengths may be capitalized upon and the weaknesses strengthened.

Aptitude Test Scores

Because there are substantial and significant individual differences in learning ability, school instruction is almost always preceded by some effort to judge the capacity of students to learn. Just as some people are taller than others, some can run faster, and some have a better ear for music, so, too, some persons learn school subjects more easily than others. Individual differences in learning ability do exist, and teachers must take these differences into account if they are to fulfill their obligation to meet the unique needs of each student.

Tests of learning ability are often called "intelligence" or "scholastic aptitude" tests. It has been well documented that scores on these tests are related to school performance. Although psychologists continue to struggle to define intelligence and to debate the nurture vs. nature issue, teachers who use scholastic ability test scores will be better served if they think of them rather narrowly and simply as indicators of future academic performance.

IQ Scores: Scholastic ability test performance is most usually reported as percentile rank scores, IQ scores, or both. The concept of the intel-

ligence quotient comes from the time when intelligence was defined as the quotient obtained by dividing a person's mental age by his chronological age. Today, IQ scores are no longer calculated in this manner, but the name persists despite rather general agreement among test experts that our schools and students would be best served if IQ scores were done away with. Scores on present IQ tests are simply standard scores with an average and spread that approximate those of the scores found on earlier IQ tests.

While the norm group for percentile rank and most other derived scores is usually made up of other students at the same grade level, the norm group for IQ scores consists of other students of the same age. This in itself diminishes the value of IQ scores for schools because the most important variable affecting what happens to a student in school is his grade placement, not his age. There are many other difficulties with IQ scores, most of which are too complex to deal with in this space. Perhaps it is sufficient to say that teachers should ignore IQs and direct their attention to percentile ranks or stanine scores whenever possible.

The greatest value of learning ability tests is that they may call attention to the few students who have unexpected discrepant scores. There are two kinds of discrepancies in which teachers should be most interested—students whose measured learning ability is quite different from their school achievement (the so-called underachievers and overachievers) and students whose abilities are very different from those of their classmates.

Learning ability test scores are rough indicators and should serve mainly as clues which stimulate further, more intensive diagnosis. Remember that low measured scholastic ability which has been substantiated by other indicators does not mean a student cannot learn. Every student can learn. Low ability means that there may be limitations to the rate of learning and the complexity of material that can be learned. Low test scores are not telling us that these students are doomed to fail. They *are* telling us, however, that they will surely fail *unless they are treated differently from average students.* By the same token, students with extremely high scholastic ability may very likely become disenchanted with schooling and either withdraw or become discipline problems *unless they are treated differently from average students.*

What I Have Not Talked About

In an effort to keep this booklet short, I have not included a number of other possible topics such as *scoring* (because today most standardized tests are machine-scored), *test administration* (because most instructions for administration furnished with testing materials cover these procedures precisely for each test), and *statistical concepts and defini-*

tions (because adequate explanation would take too much space and because such knowledge is not essential to good use of test scores by teachers). There are many excellent books and articles on these and other aspects of tests and test interpretation; some of which are included in the list of suggested readings on the following page. Finally, teachers are urged to talk with the person in the school who is responsible for testing to learn more about test interpretation and about their own school's tests.

Suggested Readings

The following two books provide readable and extensive coverage of the use and interpretation of standardized tests and of the construction of classroom tests:

Gronlund, N.E. *Measurement and evaluation in the classroom*. 2nd Edition, New York: The Macmillan Co., 1971.

Mehrens, W.A., & Lehmann, I.J. *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston, 1973.

The following publications are part of the Measurement in Education series of the National Council on Measurement in Education. These short (8-10 pages) monographs are concerned with the practical implications of educational measurement, emphasizing uses of measurement rather than technical or theoretical issues. They are available at 35 cents each from:

Office of Evaluation Service
Michigan State University
East Lansing, Michigan 48823

Aircesian, P.W., & Madaus, G.F. *Criterion-referenced testing in the classroom*. Vol. 3, No. 1.

Coffman, W.E. *On the reliability ratings of essay examinations*. Vol. 3, No. 2.

Cureton, L.W. *The history of grading practices*. Vol. 2, No. 4.

Ebel, R.L. *Shall we get rid of grades?* Vol. 5, No. 4.

Gardner, E.F. *Interpreting achievement profiles—uses and warnings*. Vol. 1, No. 2

Joselyn, E.G., & Merwin, J.C. *Using your achievement test score reports*. Vol. 3, No. 1.

Mayo, S.T. *Mastery learning and mastery testing*. Vol. 1, No. 3.

Tyler, R. *Assessing educational achievement in the affective domain*. Vol. 4, No. 3.

Warrington, W.G. *An item analysis service for teachers*. Vol. 3, No. 2.