DOCUMENT RESUME

ED 040 832 RE 002 914

AUTHOR Glock, Marvin D.
TITLE How the Classroom Teacher can use a Knowledge of Tests and Measurements.
PUB DATE 9 May 70
NOTE 15p.; Paper presented at the International Reading Association conference, Anaheim, California, May 6-9, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.85
DESCRIPTORS Achievement Gains, *Measurement, Measurement Techniques, Questioning Techniques, *Reading Research, *Reading Tests, *Teachers, Test Interpretation, Test Reliability, Test Validity

ABSTRACT

Three basic concerns in measurement were selected, and their importance for the classroom teacher were illustrated. These were test validity, reliability, and problems in measuring achievement gains. Test validity was dependent upon content, type and quality of the questions, adequacy with which the test sampled reading skills, and the care with which the test was administered. Consistency in test reliability was dependent upon the number of samples of a pupil's performance on a task and upon accurate scoring. Measuring gains in pupil achievement was dependent upon correct interpretation and treatment of scores, taking the regression effect and error factor into consideration. References are included. (CL)

ED040832

Dr. Marvin D. Glock
Cornell University
Ithaca, New York

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

How the Classroom Teacher can use a Knowledge of Tests and Measurements

Saturday, May 9, 1970
10:30-12:00 noon

Of course, there are many facets of measurement that a teacher must understand if she is to use tests effectively. Numerous books have been written about them but we are allotted only a few pages for our discussion. To discuss this topic thoroughly would require much more space than our stringent limitations allow. There are a number of alternatives open to us. Two of these are quite different. One would be to list a number of principles. The other would be to select two or three basic concerns in measurement and illustrate their importance for classroom teachers. I have chosen the latter alternative in the belief that it is more likely to be of greater practical value.

Our discussion will be limited to test validity, reliability, and the problems in measuring gains in achievement.

RE002 214

Validity

Does the test measure what it is intended to measure? One would be unwise to assume that the name or title of a test tells us what it measures. If you examine various tests of "Reading Comprehension" you will find that some require the pupil to determine the main idea of a paragraph; others demand only the retrieval of literal meaning; a few ask the reader to discern the intent and mood of the author. Tests of reading rate vary from only 60 seconds to much longer times of reading. Some give credit for speed even if many questions over the selection are not answered correctly; others give no credit for rate unless comprehension is assured. Some vocabulary tests are constructed of items listing a word with several possible synonyms from which to select the correct answer. In others, the examiner reads a sentence and the pupil is required to select one of several words to complete the sentence. In still a third type, the pupil is presented with a sentence and an underlined word. He responds by marking one of several possible synonyms. There are also vocabulary tests requiring a pupil to define a word and to use it in a sentence. It is very possible that some individuals will perform better on one of the tests of comprehension, rate, or vocabulary than on the others. Yet, all may have identical titles. Which tests are valid? Which tests are measuring what you as a teacher want them to measure?

In comprehension tests we find that some instruments are very limited in what they measure. Davis (1) lists eight skills that determine good reading comprehension: (1) recalling word meanings. (2) drawing inferences about the meaning of a word

from context. (3) finding answers to questions answered explicitly or in paraphrase. (4) weaving together the ideas in the content. (5) recognizing a writer's purpose, attitude, tone, and mood. (6) identifying a writer's techniques. (7) following the structure of a passage and (8) drawing inferences from content. Naturally, as Davis suggests, some of these skills are more important than others. However, after careful study and analysis of a particular test we found questions on only two skills, recalling word meanings and finding answers to questions answered explicitly in the passage. We would rightly conclude that this test had low content validity. It would not be a valid test for our purpose of measuring total comprehension.

A number of publishers provide an analysis chart for their tests that indicate what they believe to be the content, type of material, or skill being tested by each item. This is helpful. The procedure cannot, however, replace the need for the teacher to take the test himself - to expose himself to the tasks presented. He can then check the chart against his own judgment. It is obvious that different types of questions require pupils to respond in different ways and this, along with content, determine what the test is measuring.

There is another factor that determines what the test is measuring. We refer to the care with which the questions have been constructed and tested. A poor question may enable a pupil to select the correct answer; for example, not by determining the main idea of a paragraph but by matching a word in a question choice with the identical word in the passage.

Going to the park one day on his way to school, Bill stopped and watched them paint the new pavillion. The bright yellow color sparkled in the sunlight.

What color was the pavillion painted?

1. Red
2. Yellow
3. White
4. Green

Poor questions also allow the pupil to eliminate implausible answers and select the correct one without the comprehension intended by the author. For example, one test includes the following item to be answered after reading a selection.

The chief factor limiting the amount of land for cultivation is:
1. rugged peaks
2. climate
3. irregular coast line
4. poor farming methods

The pupil does not have to understand the passage to answer the question correctly; he could even choose the correct answer without reading the passage. Common sense dictates that none of the last three choices limits the amount of land for tillage, only the first choice could possibly be correct. Poorly constructed tests allowing test-wise pupils to respond correctly by means of irrelevant cues are not measuring what was intended.

Another factor bearing on test validity is the adequacy with which it samples reading skills or knowledge in vocabulary tests. Tests are limited in the number of responses they can ask a pupil to make. For example, in a vocabulary test only a few of the words that a pupil might be expected to know are included. Another vocabulary test might present an entirely different list of words. The manner in which these lists are

selected will determine how well the test depicts vocabulary development. Some lists include general vocabulary; others may be loaded with scientific terminology. Various kinds of bias can exist.

The care with which the test is administered can influence validity. Scores may be consistently too high or too low because of administrative procedures. In standardized tests instructions must be carefully followed giving no more nor less help than is specified. Time limits must be adhered to. Room conditions and seating arrangements should provide for optimal performance. Interruptions and other distractions must be eliminated. No teacher should administer any standardized test without first carefully reading through all instructions and underlining the time limits, in color preferably. It is assumed, of course, that all pupils will answer the same questions.

Reliability

Another important quality of a good test is adequate reliability. No psychological test can measure as precisely as a foot-rule or even a house-hold scale. On the other hand, good tests do reflect the quality of reliability. Reliability to describe tests has a different connotation than when the word is used in common parlance. When we speak of a reliable person, we imply veracity and complete dependability on what the person does and says. On the other hand, a reliable test implies that it is consistent in what it does measure. A very reliable test may not be telling the truth, but it continues to report the same falsehood quite accurately. For example, if we administer a paper and pencil verbal intelligence test to

a child who cannot read, he would invariably make a low score, an indication of low intelligence. The test might have high reliability, but it would not be measuring intelligence. Rather, it would be revealing a child's inability to read. It would not be telling us the truth about his intelligence.

Let's use another illustration to help us understand the concept of reliability. Suppose we measure your height with a yardstick that was marked off only in feet. There are only two marks on the stick. It is obvious that the measurement of your height would probably be less accurate than if the stick was calibrated to 1/16 of an inch. Getting two measurements alike with the rough markings is most unlikely. Such a poor instrument would give inconsistent measurements; it is unreliable. Likewise, if we are to have confidence in a test score, it must be attained by careful measurement.

There are certain factors that determine a test's reliability. First, we must have a number of samples of a pupil's performance on a task. We certainly don't judge a batter's skill by one time at bat. His batting average is determined over a series of performances at the plate. Neither do we judge a pupil's vocabulary effectively by giving him a word or two to define - nor even 10 or 20 words. He must give the meanings of a great many more words to get a precise measure of his ability. By the same token, we certainly won't be able to have much confidence in a reading comprehension test score unless the child answers a considerable number of comprehension questions. The more questions; that is, the more samples, the more likely is the test to be reliable.

That's the reason part scores on a test are so often suspect. Each part is a small test; only a few questions yield a part score. Sampling with such a limited number of tasks is inadequate to give a reliable score.

A test's reliability also depends upon accurate scoring. Scoring is not difficult with objective type tests; there is an obvious right or wrong answer. It becomes more of a problem when pupils write out their answers to a question. There are specific procedures to improve scoring reliability in these instances.

Of course, there is available data provided with all good tests for the user to determine the adequacy of the test's reliability. In general, for an individual pupil's test score to be reliable enough for proper interpretation, there should be at least one half hour of testing time with a minimum of 40 to 50 questions. A reading test with items demanding complex and critical thinking will need more questions for optimum reliability than does a test requiring mastery of literal meaning and factual information. Insufficient length prevents subtests of ten to fifteen questions from attaining adequate reliability. Also a test that is designed for testing pupils in a wide-range of grades, e.g. 3 through grade 12, may have only a few questions that are suitable - not too easy or too difficult - for the children in any one grade. This results in a very short test for each pupil.

We have been discussing reliability chiefly in terms of standardized tests. How is this information related to the short, teacher-made, classroom tests? Well, certainly the

teacher doesn't want to make important decisions on the basis of test performance involving only four or five questions. However, over a period of time, if a teacher is consistent about administering these short tests, he will build up a considerable number of questions - in effect a long test - whose reliability will most likely be adequate to aid with the help of other information in making valid judgements about each pupil.

But no test is perfectly reliable. For the practical situation there is always an error of measurement in a test score. Therefore, when a pupil earns a score on a test we never know whether it is higher or lower than he deserved. Many of the better tests use a special system to help the teacher interpret test scores. Raw scores are changed to percentiles and a score is reported as falling within a percentile band. For example, in Figure 1 if Mary's converted score in reading was 300, the band between the 60th to the 82nd percentiles would be an indication of the possible error. However, we could say with reasonable certainty that her score was better than 60% to 82% of the standardization group. Also, since this is a battery of tests, we could draw some conclusions about Mary's comparative performance in several subject areas. The percentile bands of reading and writing overlap. Therefore, we could not conclude with assurance that her reading ability was greater than the score in writing because of the measurement error in both scores. However, it would be reasonable to conclude that her reading score does represent superiority when compared with listening, science, and mathematics, because there is no overlapping.

Name White Alice Jane
Last First Middle

School Pleasant Grove Grade or Class 11

Age 16 Years 3 Months Date of Testing Fall Fall or Spring 1967 Year

Norms Used

☒ Publisher's ☒ Fall Grade or Class 11th Grade
☐ Local ☐ Spring Other _____

© Copyright 1957, All rights reserved

Cooperative Test Division (ETS) Educational Testing Service · Princeton, N.J. · Los Angeles 27, Calif.

Here you can profile a student's percentile ranks on as many as six tests in the STEP series. In order for your comparisons between the areas to be valid, all tests included should be administered within a period of 4 or 5 months.

Recording. Directions for recording information and drawing percentile bands on the PROFILE form are included in each STEP MANUAL FOR INTERPRETING SCORES. Consult the manuals for the tests used.

Interpreting. To compare a student's performance on one of the tests in the STEP series with that of students in the norms group used, note the unshaded parts of the column above and below the percentile band. For example, if the Listening percentile band is 24-36, you know that 24 per cent of students in the norms group score lower than this student and 64 per cent score higher. In other words, this student's Listening performance is below average with respect to the norms group.

To compare a student's standings on any two tests in the STEP series, the following rules apply:

1. If the percentile bands for any two tests overlap, there is no important difference between the student's standings on those two tests.
2. If the percentile bands for any two tests do *not* overlap, standing represented by the higher band is really better than standing represented by the lower band.

Examples: According to local norms, a student's percentile bands for three tests are

Mathematics (2A)	50-62
Social Studies (2A)	60-71
Science (2B)	41-52

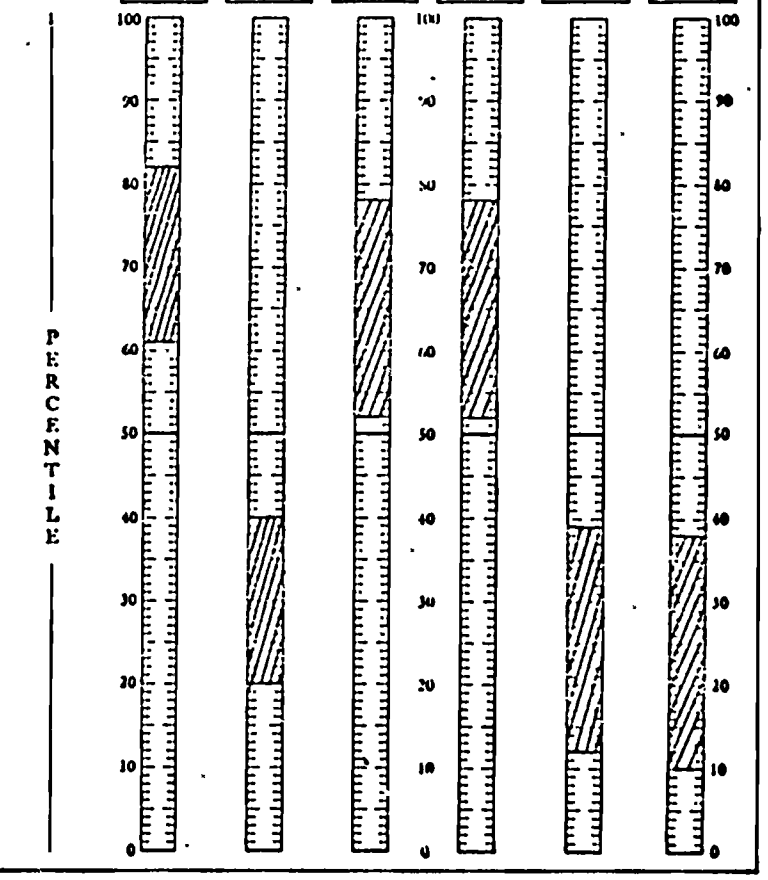
Bands for Mathematics and Social Studies overlap; there is no important difference between the student's standings in these two areas. The same is true of Mathematics and Science. However, bands for Science and Social Studies do not overlap; the student's standing in Social Studies is higher than his standing in Science. We expect the student to perform better in the social studies area than in the science area.

More detailed discussions of interpretations are contained in each STEP MANUAL FOR INTERPRETING SCORES.

D107H30V

STEP STUDENT PROFILE
SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS

Test	Reading	Listening	Writing	Soc. Studies	Science	Mathematics
Form	2A	2A	2A	2A	2A	2A
Converted Score	300	280	290	280	270	265



From Cooperative Test Division, 1957; reproduced by permission of the Educational Testing Service.

FIGURE 23 | PUPIL PROFILE CHART.

The teacher who realizes that he must interpret test scores with great care because none is free of error, will also muster as much additional information as possible before making important judgments and decisions about children.

Measuring Gains in Achievement

One purpose of tests is to determine the progress of pupils. Initial and final scores on standardized achievement tests are often determined. But when a test measuring skills such as reading, arithmetic, and writing, which develop more or less continuously, is administered at the beginning and end of the school year the average score is almost certain to rise. However, if we look at the following scores in Figure 2 we note a phenomenon that could be embarrassing. Note the difference in gains among the various groups. What do we see? We find that the lowest group has gained the most, the highest group the least and the middle groups marked gains in between. Some of the high group in other supject areas would appear to have lost achievement. How can we account for this state of affairs? Have the lowest students actually learned more because instruction was pitched at their level while in the meantime the high achievers just marked time?

This might seem reasonable if in this well-known study the same phenomenon had not occurred with interest inventories, inventories of beliefs, and problems in human relations. Even in the affective realm of attitudes, values, and personal-social adjustment those who made the lowest scores gained the most on a second administration and those making the highest score gained the least if they did not actually make a lower score.

An Analysis of Gains in Reading and Writing

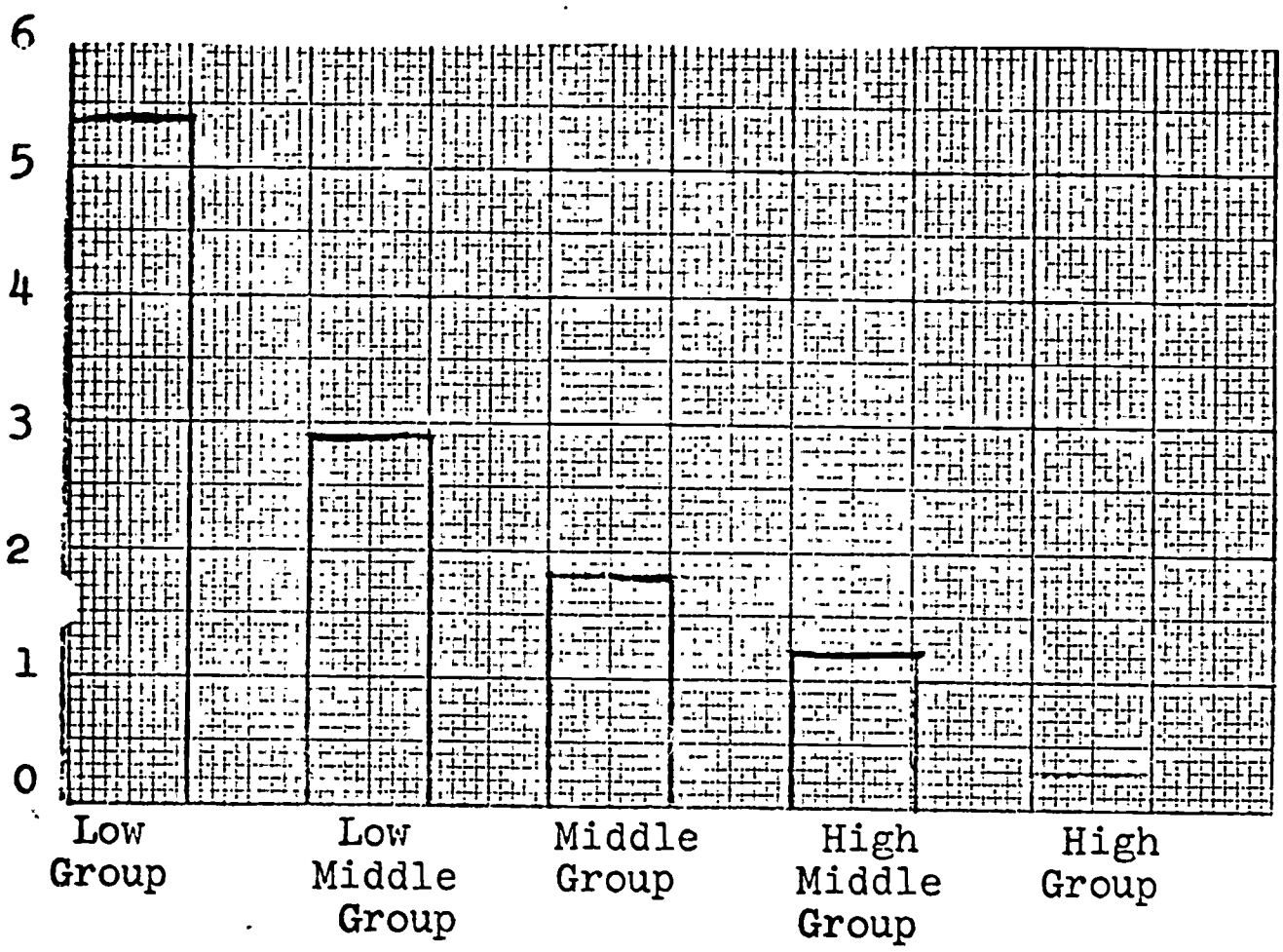


Figure 1 - Average Gains of Students on post-tests, classified according to pretest standing (2)

If we are not to conclude that students with initial high scores change their behavior very little if at all while those with low scores learn a great deal, then we must look to the construction of our tests or to the testing repetition for our answer.

One answer to the problem is that the tests may have been too easy for the better pupils so that they very nearly answered all of the questions on the first testing. There would be little if any opportunity for improvement because another form of the test would be of the same level of difficulty.

Another explanation of why lower achievers make greater test score gains than higher achievers is the phenomenon of "regression". In 1889 Galton, an Englishman, reported that short parents tended to have children who were taller than they and tall parents tended to have offspring who were shorter than they. (3) He stated that:

> However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by observation, that the stature of the adult offspring must, on the whole, be more mediocre than the stature of their parents. (4)

Galton called this tendency the law of regression and related it to various hereditary traits. It can be easily explained. There is not a perfect correlation between parents' and children's height. Therefore, the children of tall parents will be shorter and closer to the mean. If there was a perfect correlation, they would be as tall as their parents. They can't be taller since their parents are already at the extreme end of the distribution. The same kind of reasoning holds for the relationship between short parents and their children.

Regression is observable whenever we have two variables such as height and weight, scores from two achievement tests, or a score from an ability test and an achievement test, that are not perfectly correlated. Then there exists a tendency for students who make the highest scores on achievement test number one to make less superior scores on test number two. Pupils who make high scores on ability tests tend to make not so high scores on achievement tests or school marks.

What are the implications of this phenomenon for teaching? Obviously we need norms for gains as well as for status and norms based on each initial score. Teachers can develop their own norms and also they can keep records of pupils so that they can interpret gains more validly. For example, it may be that a gain of 40% more correct items from a low score may be generally expected while a gain of 20% more correct from a higher score may represent an exceptional improvement. In comparing gains we must be aware of the initial score. Was it low or high?

One other reason why we should exercise care in the use of gain scores is the error factor. The initial and final test measurements each contain error. Errors then accumulate when the scores are subtracted to determine the gain. The difference score may, therefore, be a representation of error rather than gain. Seldom are tests available to measure reliable short term gains for individual pupils. It is possible, however, to use the difference between means of initial and final testing to determine the effectiveness of instruction for a class.

In summary, a valid test for a teacher's purpose measures what he wants it to measure. He cannot depend on the test's title for this assurance. He must search further and examine the items in addition to reading the information provided in test manuals. A valid test must also be a reliable test. On the other hand, a reliable test isn't necessarily a valid one because it needs only measure consistently what it does measure. Reliability is a necessary but not a sufficient condition for a good test. However, no test is perfectly reliable. Each contains some error and it is important that we are cognizant of this fact when making decisions about children.

Gains on test scores must be interpreted with considerable care because of the regression effect. Gains have different meanings for initial low and high scores. It has been suggested that norms be provided that are based on the magnitude of these initial scores.

References

1. Davis, F. B. Identification and measurement of reading skills of high school students. Washington, D. C.: Office of Education, U. S. Department of Health, Education, and Welfare, 1967.

2. Dressel, Paul L. & Mayhew, Lewis B. General education: exploration in evaluation. Washington: American Council on Education, 1954. Pp. 59, 99, 128, 166, 204, 227, and 237.

3. Galton, Francis Natural Inheritance. London: Macmillan, 1889, p. 95.

4. Ibid, p. 106.