DOCUMENT RESUME

ED 079 385 TM 002 985

AUTHOR Wesman, Alexander G.
TITLE Comparability Vs. Equivalence of Test Scores.
INSTITUTION Psychological Corp., New York, N.Y.
REPORT NO Bull-53
PUB DATE Sep 58
NOTE 4p.; Reprint from Test Service Bulletin
JOURNAL CIT Test Service Bulletin; n53 p6-9 Sep 1958

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Bulletins; Norms; *Scores; Standardized Tests; *Test Interpretation; *Test Results

ABSTRACT

Comparable scores represent equal rank in a given population-they imply nothing concerning what is being measured. Equivalent scores represent similarity of what is being measured -- the more complete the equivalence, the greater the likeness of measurement. The test user would do well to keep these distinctions in mind; to avoid being confused into accepting comparable scores as denoting equivalence of measurement; to insist on close approximations to complete equivalence in alternate forms; and to recognize that in substituting one test for another he may prefer rough equivalence to more precise equivalence if he is seeking to improve validity. (Author)

THE PSYCHOLOGICAL CORPORATION INCORPORATED

Test Service Bulletin

No. 53 THE PSYCHOLOGICAL CORPORATION September, 1958

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to 304 East 45th Street, New York 17, N. Y.

COMPARABILITY VS. EQUIVALENCE OF TEST SCORES

THE vagaries of the English language must be a source of considerable bewilderment to those who are faced suddenly with the need to learn our tongue. How does one learn what "fix" means? The mariner who wishes to determine his position gets a fix; the professional crook seeks a game that he can fix; the squeaky door needs to be fixed; a committee chairman fixes a date; a student eyes his professor with a fixed stare; a culprit is found and blame is fixed—and the culprit finds himself in a fix. So, too, the special language of tests and measurements contains some ambiguities. But lest they interfere with clear understanding of important concepts, ambiguities ought to be clarified. Two of the common words in the testing field which are surrounded by confusion are "comparable" and "equivalent."

Two test scores are *equivalent* if either can properly be substituted for the other. Both the trait being measured and the means of measurement must correspond. Two scores may be *comparable*, on the other hand, yet reflect very dissimilar abilities. In fact, the scores may be numerically quite different — may even be expressed in different units — and still be comparable.

Comparability properly refers merely to rank in a group; the term carries no connotation with respect to what is being measured. For example, within the *Differential Aptitude Tests* a score of 56 on the Clerical Speed and Accuracy test is comparable for certain individuals to a score of 47 on the Mechanical Reasoning test. In each case, the score represents the 70th percentile for tenth grade boys in the population used in standardizing the tests. The central fact to be noted is that *two scores are comparable if they represent the same standing in the same population.* There is no implication that the scores denote the same, or even similar, abilities. Even casual inspection of the two tests reveals how little they measure in common. In fact, the average correlation between the Mechanical and Clerical tests is about .10.

If a low coefficient of correlation between two sets of test scores doesn't preclude comparability, neither does a high one assure it. As indicated above, the size of the correlation coefficient is irrelevant to the matter of comparability. Scores on the *DAT* Numerical Ability test and the *Stanford* Arithmetic test are not comparable even though these two tests may be expected to correlate about .75. Scores on these tests are not comparable because the tests were not standardized on the same population. For similar reasons, scores on the *DAT* Space Relations test are not comparable to scores on the *Revised Minnesota Paper Form Board*, although both are tests of space perception. It is not what the tests measure but the population used in standardizing the tests that determines comparability.

But, we may ask, if comparability is merely a matter of giving two tests to one population, cannot one make *any* two tests comparable by giving them to a single group? Yes, indeed. Any school or business organization *can* develop sets of comparable scores by giving any two (or more) tests to its students or employees.

Will such data then be useful to other institutions? That depends on the resemblance between the group on which comparability of scores was based, and the group with which the result is to be used. If the groups are sufficiently alike, a table of comparable scores will apply about as well to the second as it does to the first group. If, on the other hand, the two groups are unlike in some important respect (e.g., age, sex, education, relevant environment, etc.), it may be inadvisable to assume that the table of comparable scores will apply as well to the second group. For example, among tenth grade boys a score of 44 on the *DAT* Mechanical Reasoning test is comparable to a score of 34 on the *DAT* Sentences test; both are at the sixtieth percentile for this norms group. Among tenth grade girls, the same Mechanical Reasoning score of 44 is comparable to a Sentences score of 66; both are at the ninety-fifth percentile for girls in the tenth grade. Like norms and validity, comparability is specific to the group on which the data are obtained.

It may seem surprising that two scores which represent equal standing in one group may reflect quite different standings in another group. Some thoughtful consideration, however, will make it evident that such variations in comparability should be expected. An example may help to illuminate the issue. Let us suppose that a test of English grammar and a test of reading comprehension in French have been administered to two groups of students. Group A consists of freshmen who have had only three months of exposure to the learning of French; Group B consists of sophomores who have just completed two years of course work in the subject. We now prepare distributions of scores for the pair of tests and then compute percentiles to show what per cent of students fall below each score on each test. We compute these percentiles separately for the freshmen and sophomores. For the freshmen, we find the score at the 50th percentile on the English grammar test, and the score at the 50th percentile on the French reading comprehension test. These two scores are comparable — *for the freshmen*. What happens when we seek similarly comparable scores for the sophomores? On the English grammar test the score which is at the 50th percentile for sophomores is likely to be a little higher than the median for freshmen. The French comprehension score at the 50th percentile for sophomores is likely to be very much higher than the score at the 50th percentile for freshmen. The increased knowledge of French represented by the additional year and two-thirds of study will have a far greater effect on the French test scores than an additional year of exposure to English. We may expect, then, that the French score comparable to a particular score in English will be appreciably higher for sophomores than for freshmen.

Table I has been prepared to illustrate the situation. Inspection of the table shows that, for freshmen who have studied French for three months, an English grammar score of 58 is comparable to a French comprehension score of 64. For sophomores who have finished two years of French, however, an English grammar score of 58 is comparable to a French comprehension score of 73—a substantial difference. Clearly, any attempt to apply these freshman data on comparability to the sophomores would result in serious error. Proper interpretation of comparable scores requires that we know the characteristics of the group on which comparability was established. If we wish to apply published tables of comparable scores to our local population, we need to assure ourselves that the groups are sufficiently similar to permit such generalization.

Perhaps the most important distinction between "comparability" and "equivalence" is that, *whereas test content is irrelevant to comparability, test content is fundamental to equivalence*. Two test scores are equivalent if they can properly be substituted for one another. Essentially, this means that scores from one test must represent

TABLE I. Illustrative Norms for Two Groups.

Percentile	Freshmen		Sophomores	
	English Grammar	French Reading	English Grammar	French Reading
99	82	88	82	98
97	78	85	79	95
95	75	82	76	92
90	72	79	74	89
85	70	77	72	86
80	68	75	70	84
75	67	73	68	83
70	65	72	67	81
65	64	70	66	80
60	63	69	64	79
55	62	68	63	78
50	61	67	62	76
45	59	66	61	75
40	58	64	60	74
35	57	63	58	73
30	56	62	57	72
25	55	60	56	70
20	53	59	54	68
15	51	57	52	66
10	49	54	50	64
5	46	51	47	61
3	44	48	45	58
1	40	44	42	54

This table is but slightly adapted from tables of norms found in the published manuals for a test of English grammar and a test of French reading comprehension. The scores are standard scores based on a single scale, with a standard deviation of approximately 10.

the same psychological or educational qualities in the individual as do scores from the other test. Most precisely, two tests are completely equivalent if their content is essentially identical and they measure with equal precision (reliability). If these conditions are met, it does not matter which of the two tests is used. These conditions are ordinarily most closely approximated where parallel forms of a test have been constructed – forms which are intended to be interchangeable.

When parallel forms of a test are available, there is ordinarily the implicit, if not explicit, assumption that these forms *are* actually interchangeable. This means that we have no basis for suggesting that a person take one form rather than another—the information obtained will be of equal value whether Form A or Form B is administered. The specific items in one form are of no greater significance than the items which happen to be in the alternate form.

Assumptions we can make with regard to content and reliability of parallel forms of one test are not readily acceptable when we are dealing with two somewhat different tests of the same general ability. This situation is one in which the problem of equivalence frequently arises. For example, a counselor may have reading comprehension scores from the *Stanford Achievement Test* for some pupils, and scores from the *Iowa Silent Reading Test* for other pupils; or, an industrial organization may wish to substitute a modern clerical aptitude test for an outmoded one. In such cases, it is important to know the degree of equivalence of the scores from the two reading tests, or the two clerical tests.

In these circumstances, the size of the coefficient of correlation between the tests is of prime importance. Obviously, lack of perfect reliability in each of the tests will prevent the correlation coefficient from reaching 1.00. Even disregarding the effects of unreliability, however, the correlation would still be less than perfect because each reading test was constructed somewhat differently from the other; the two clerical tests were also prepared according to distinctly different plans. The greater the divergence in specific abilities measured, the more ambiguous the term "equivalent" becomes.

If the correlation coefficient is 1.00, we can say with complete confidence that all persons who score in, say, the sixth decile (51st to 60th percentiles) on one test will also score in the sixth decile on the other. If the coefficient is .90, we may expect that, of those who score in the sixth decile on one test, 22.5% will score in the sixth decile on the other test; the remaining 77.5% will be distributed as follows: approximately 20% each in the fifth and seventh deciles, about 13% in the fourth and eighth deciles, and the remainder in the second, third, ninth, and tenth deciles. If the coefficient is .75, of those who score in the sixth decile on the first test, we may expect 15.2% to score in the sixth decile on the second test. The other examinees would be found in the first decile (1.9%), the second decile (6.0%), the third (9.6%), the fourth (12.4%), the fifth (14.2%), the seventh (14.6%), the eighth (12.9%), the ninth (9.4%), and the tenth decile (3.8%). In these circumstances, we cannot say that an individual will certainly achieve the same score on one test as he does on another. Instead, we can speak only of the *probability* that people who make a certain score on one test will obtain various scores on the other.*

In practice it would be extremely awkward to present a table of equivalents in terms of these probabilities. To simplify matters, we present pairs of individual scores as equivalents – usually based on the equi-percentile method or a variant of it. That is, we find for a given group those scores which are at the 40th percentile on forms A and B, and present those scores as equivalent. What distinguishes the procedure from that in which we obtain comparability of scores is that we have in the equivalence table the assumption that what is being measured is the same in the two forms.

Does this mean that an older test cannot be replaced by a newer and presumably better test? Not at all. To persist in the use of instruments when more valid or more efficient tests become available is poor practice. The heart of any test use is validity—whether the test is doing what it is intended to do. If test N can offer appreciably better prediction than test Q, test N should replace test Q in the particular situation; in this case we do not want a truly equivalent test—we want a better test. If we have had a good deal of experience with test Q, we may wish to know the relative rank represented by specific scores on tests Q and N. If we have used a cutoff score on test Q, we may wish to know what score on test N would eliminate a similar proportion of the applicants. This information can be obtained by giving both tests to the same population, or to two very similar populations.

The resulting table of matched scores is a table of comparability. To evaluate the degree to which the table is also a table of equivalents, we need to know the coefficient of correlation between the two sets of scores. If the scores are comparable and we use the same cutoff score on test N that we used for test Q, we will accept the same number of applicants. Because the tests are not perfectly

*The above statements apply to alternate forms of tests as well as to tests intended to measure somewhat different abilities. If alternate forms of a test correlate .75, the per cents to be expected in each decile will be the same as for a coefficient of .75 between non-parallel tests.

reliable not precisely equivalent, we will not accept precisely the same individuals by means of the two tests--and because test N is more valid, we will accept a larger number of good applicants and a smaller number of prospective failures. This is an outcome much to be desired. We are obviously not seeking precise equivalence. We are happy to trade some precision in the equivalence for some improvement in validity.

To summarize, comparable scores represent equal rank in a given population—they imply nothing concerning what is being measured. Equivalent scores represent similarity of what is being measured -- the more complete the equivalence, the greater the likeness of measurement. The test user would do well to keep these distinctions in mind; to avoid being confused into accepting comparable scores as denoting equivalence of measurement; to insist on close approximations to complete equivalence in alternate forms; and to recognize that in substituting one test for another he may prefer rough equivalence to more precise equivalence if he is seeking to improve validity.--A.G.W.

A New Reading Test for Use in High Schools and Colleges

• DAVIS READING TEST

Frederick B. Davis and Charlotte Croon Davis

Carefully constructed to measure the reading skills of college freshmen and high school juniors and seniors, this new reading test provides scores in:

1. Level of comprehension
2. Speed of comprehension

The Level score indicates the depth of understanding displayed by a student in reading the kinds of material he is ordinarily required to read in high school and college; the Speed score indicates the rapidity and accuracy with which he understands the same material.

Passages varying in length from five to thirty lines are used as a basis for multiple-choice items measuring five categories of reading skills:

- ▶Finding the answers to questions answered (explicitly or in paraphrase) in a passage;
- ▶Weaving together the ideas in a passage and grasping its central thought;
- ▶Making inferences about the subject of a passage and about its author's purpose or viewpoint;
- ▶Recognizing the tone and mood of a passage and the literary devices used by its author;
- ▶Following the structure of a passage, as in identifying antecedents and referents.

The *Davis Reading Test* is available in four equivalent forms. In each form of the test, the first and second halves have been carefully equated. Within the 40-minute time limit nearly every examinee completes the first half, and the Level of Comprehension score is based on this portion. The Speed of Comprehension score is based on the whole test.

The test may be scored quickly and easily either by machine or by hand. Raw scores are converted into scaled scores representing the same relative amounts of ability in either Level of Comprehension or Speed of Comprehension, regardless of the form of the test used. Percentile norms are provided for students in the eleventh and twelfth grades and for college freshmen. The standardization is based on over 18,000 students in 18 colleges and 29 high schools in 25 states.

Marked discrepancies between scaled scores for Level and for Speed, or scores which are markedly low for the student's grade, indicate a need for individual diagnosis and remedial reading help. Further information of diagnostic value may be obtained by comparing *Davis Reading Test* percentiles with results on the *College Qualification Tests.*

In 1963, Series 2 became available for grades 8-11, to supplement Series 1 at the grade 11 to-college level. There are four forms at each level:

Grades 8-11: Forms 2A, 2B, 2C, 2D
Grades 11-13: Forms 1A, 1B, 1C, 1D

For packaging and prices of the test booklets, answer sheets, and accessories, see the Test Catalog.