DOCUMENT RESUME

ED 079 347 TM 002 946

AUTHOR	Seashore, Harold G.
TITLE	Methods of Expressing Test Scores.
INSTITUTION	Psychological Corp., New York, N.Y.
REPORT NO	Bull-48
PUB DATE	Jan 55
NOTE	4p.; Reprint from Test Service Bulletin
JOURNAL CIT	Test Service Bulletin; n48 p7-10 Jan 1955
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bulletins; *Intelligence Tests; *Raw Scores; *Scores; *Statistical Analysis; *Test Results

ABSTRACT

The simplicity of standard score syst s, percentile equivalents, and their relation to the ideal normal di tribution are discussed and illustrated. Standard scores are z-scores, the T-scores, College Entrance Examination Board scores, and Army General Classification Test scores. A derivative of the general standard score system is the stanine plan, which divides the norm population into nine groups and nine percentages which indicates the percent of the total population in each of the stanines. Interpretation of the Wechsler scales depends on a knowledge of standard scores. A subject's raw score on each of the subtests in these scales is converted, by appropriate norm tables, to a standard score, based on a mean of 10 and a standard deviation of 3. The sums of standard scores on the Verbal Scale, the Performance Scale, and the Full Scale are then converted into IQ's. These IQ's based on a standard score mean of 100; the standard deviatn of the IQ's set at 15 points. IQ's of the type used in the Wechsler scales are known as "deviation IQ's" as contrasted with the IQ's developed from scales in which a derived mental age is divided by chronological age. (For related document, see TM 002 944.) (DB)

Test Service Bulletin

No. 48 THE PSYCHOLOGICAL CORPORATION January, 1955

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.

METHODS OF EXPRESSING TEST SCORES

AN individual's test score acquires meaning when it can be compared with the scores of well-identified groups of people. Manuals for tests provide tables of norms to make it easy to compare individuals and groups. Several systems for deriving more meaningful "standard scores" from raw scores have been widely adopted. All of them reveal the relative status of individuals within a group.

The fundamental equivalence of the most popular standard score systems is illustrated in the chart on the next page. We hope the chart and the accompanying description will be useful to counselors, personnel officers, clinical diagnosticians and others in helping them to show the uninitiated the essential simplicity of standard score systems, percentile equivalents, and their relation to the ideal normal distribution.

Sooner or later, every textbook discussion of test scores introduces the bell-shaped normal curve. The student of testing soon learns that many of the methods of deriving meaningful scores are anchored to the dimensions and characteristics of this curve. And he learns by observation of actual test score distributions that the ideal mathematical curve is a reasonably good approximation of many practical cases. He learns to use the standardized properties of the ideal curve as a model.

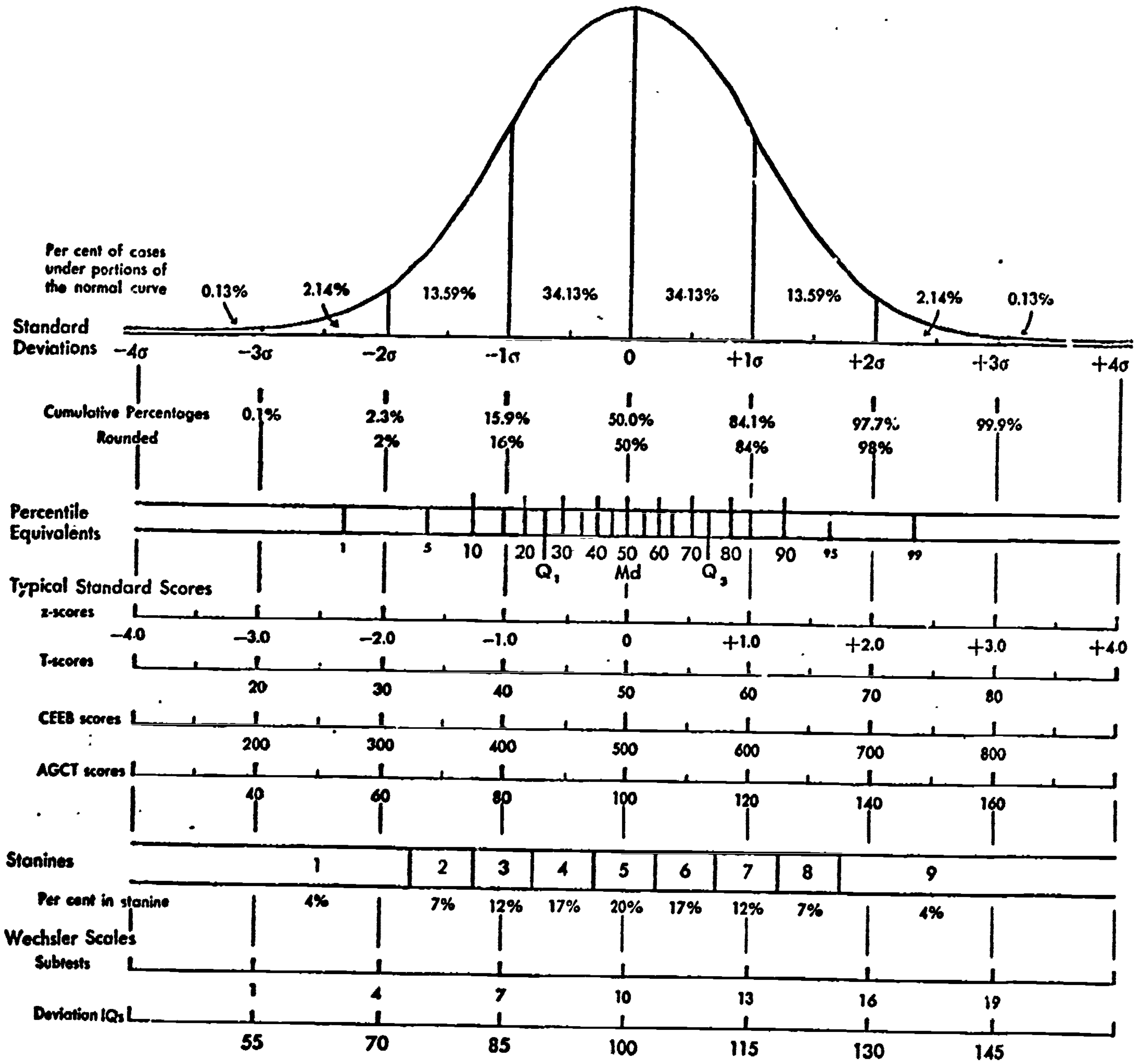
Let us look first at the curve itself. Notice that there are no raw scores printed along the baseline. The graph is generalized; it describes an idealized distribution of scores of any group on any test. We are free to use any numerical scale we like. For any particular set of scores, we can be arbitrary and call the average score zero. In technical terms we "equate" the mean raw score to zero. Similarly we can choose any convenient number, say 1.00, to represent the scale distance of one standard deviation.[1] Thus, if a distribution of scores on a particular test has a mean of 36 and a standard deviation of 4, the zero point on the baseline of our curve would be equivalent to an original score of 36; one unit to the right, $+1\sigma$, would be equivalent to 40, $(36 + 4)$; and one unit to the left, -1σ, would be equivalent to 32, $(36 - 4)$.

The total area under the curve represents the total number of scores in the distribution. Vertical lines have been drawn through the score scale (the baseline) at zero and at 1, 2, 3, and 4 sigma units to the right and left. These lines mark off subareas of the total area under the curve. The numbers printed in these subareas are per cents—*percentages of the total number of people.* Thus, 34.13 per cent of all cases in a normal distribution have scores falling between 0 and -1σ. For practical purposes we rarely need to deal with standard deviation units below -3 or above $+3$; the percentage of cases with scores beyond $\pm 3\sigma$ is negligible.

The fact that 68.26 per cent fall between $\pm 1\sigma$ gives rise to the common statement that in a normal distribution roughly two-thirds of all cases lie between plus and minus one sigma. This is a rule of thumb every test user should keep in mind. It is very near to the theoretical value and is a useful approximation.

Below the row of deviations expressed in sigma units is a row of per cents; these show *cumulatively* the percentage of people which is included *to the left* of each of the sigma points. Thus, starting from the left, when we reach the line erected above -2σ,

[1]The mathematical symbol for the standard deviation is the lower case Greek letter sigma or σ. These terms are used interchangeably in this article.



NOTE: This chart cannot be used to equate scores on one test to scores on another test. For example, both 600 on the CEEB and 120 on the AGCT are one standard deviation above their respective means, but they do not represent "equal" standings because the scores were obtained from different groups.

we have included the lowest 2.3 per cent of cases. These percentages have been rounded in the next row.

Note some other relationships: the area between the $\pm 1\sigma$ points includes the scores which lie above the 16th percentile (-1σ) and below the 84th percentile ($+1\sigma$) — two major reference points all test users should know. When we find that an individual has a score 1σ above the mean, we conclude that his score ranks at the 84th percentile in the group of persons on whom the test was normed. (This conclusion is good provided we also add this clause, at least subvocally: *if this particular group reasonably approximates the ideal normal model.*)

The simplest facts to memorize about the normal distribution and the relation of the *percentile* system to deviations from the average in sigma units are seen in the chart. They are

Deviation from the mean	-2σ	-1σ	0	$+1\sigma$	$+2\sigma$
Percentile equivalent	2	16	50	84	98

To avoid cluttering the graph reference lines have not been drawn, but we could mark off ten per cent sections of area under the normal curve by drawing lines vertically from the indicated decile points (10, 20, . . . 80, 90) up through the graph. The reader might do this lightly with a colored pencil.

We can readily see that ten per cent of the area (people) at the middle of the distribution embraces a smaller *distance* on the baseline of the curve than ten per cent of the area (people) at the ends of the range of scores, for the simple reason that the curve is much higher at the middle. A person who is at the 95th percentile is farther away from a person at the 85th percentile in units of *test score* than a person at the 55th percentile is from one at the 45th percentile.

The remainder of the chart, that is the several scoring scales drawn parallel to the baseline, illustrates variations of the *deviation score* principle. As a class these are called *standard scores*.

First, there are the *z-scores*. These are the same *numbers* as shown on the baseline of the graph; the only difference is that the expression, σ, has been omitted. These scores run, in practical terms, from −3.0 to +3.0. One can compute them to more decimal places if one wishes, although computing to a single decimal place is usually sufficient. One can compute z-scores by equating the mean to 0.00 and the standard deviation to 1.00 for a distribution of any shape, but the relationships shown in this figure between the z-score equivalents of raw scores and percentile equivalents of raw scores are correct only for normal distributions. The interpretation of standard score systems derives from the idea of using the normal curve as a model.

As can be seen, T-scores are directly related to z-scores. The mean of the raw scores is equated to 50, and the standard deviation of the raw scores is equated to 10. Thus a z-score of +1.5 means the same as a T-score of 65. T-scores are usually expressed in whole numbers from about 20 to 80. The T-score plan eliminates negative numbers and thus facilitates many computations.[2]

The College Entrance Examination Board uses a plan in which both decimals and negative numbers are avoided by setting the arbitrary mean at 500 points and the arbitrary sigma at another convenient unit, namely, 100 points. The experienced tester or counselor who hears of a College Board SAT-V score of 550 at once thinks, "Half a sigma (50 points) above average (500 points) on the CEEB basic norms." And when he hears of a score of 725 on SAT-N, he can interpret, "Plus 2¼σ. Therefore, better than the 98th percentile."

[2]T-scores and percentiles both have 50 as the main reference point, an occasional source of confusion to those who do not insist on careful labelling of data and of scores of individuals in their records.

During World War II the Navy used the T-score plan of reporting test status. The Army used still another system with a mean of 100 and a standard deviation of 20 points.

Another derivative of the general standard score system is the *stanine* plan developed by psychologists in the Air Force during the war. The plan divides the norm population into nine groups, hence, "standard nines." Except for stanine 9, the top, and stanine 1, the bottom, these groups are spaced in half-sigma units. Thus, stanine 5 is defined as including the people who are within $\pm 0.25\sigma$ of the mean. Stanine 6 is the group defined by the half-sigma distance on the baseline between $+0.25\sigma$ and $+0.75\sigma$. Stanines 1 and 9 include all persons who are below -1.75σ and above $+1.75\sigma$, respectively. The result is a distribution in which the mean is 5.0 and the standard deviation is 2.0.

Just below the line showing the demarcation of the nine groups in the stanine system there is a row of percentages which indicates the per cent of the total population in each of the stanines. Thus 7 per cent of the population will be in stanine 2, and 20 per cent in the middle group, stanine 5.

Interpretation of the Wechsler scales (W-B I, W-B II, WISC, and WAIS) depends on a knowledge of standard scores. A subject's raw score *on each of the subtests* in these scales is converted, by appropriate norms tables, to a standard score, based on a mean of 10 and a standard deviation of 3. The sums of standard scores on the Verbal Scale, the Performance Scale, and the Full Scale are then converted into IQs. These IQs are based on a standard score mean of 100, the conventional number for representing the IQ of the average person in a given age group. The standard deviation of the IQs is set at 15 points. In practical terms, then, roughly two-thirds of the IQs are between 85 and 115, that is, $\pm 1\sigma$.[3] IQs of the type used

[3]Every once in a while we receive a letter from someone who suggests that the Wechsler scales ought to generate a wider range of IQs. The reply is very simple. If we want a wider range of IQs all we have to do is to choose a *larger arbitrary* standard deviation, say, 20 or 25. Under the present system, $\pm 3\sigma$ gives IQs of 55 to 145, with a few rare cases below and a few rare cases above. If we used 20 as the standard deviation, we would *arbitrarily* increase the $\pm 3\sigma$ range of IQs from 55-145 to 40-160. This is a wider range of numbers! But, test users should never forget that adaptations of this kind do not change the responses of the people who took the test, do not change the order of the persons in relation to each other, and do not change the psychological meaning attached to an IQ.