A Description of Scores from Standardized Tests

Standardized instruments yield a variety of scores, most of which are derived from raw scores (usually number correct scores). There are two broad classes of scores, including norm-referenced scores and criterion-referenced scores.

Criterion-referenced scores are those that indicate how much an individual knows or what an individual can do. This can only be accomplished if the content domain is clearly defined so that we can refer to the content domain from the score itself. For example, a criterion-referenced interpretation would be where we say: If you get 85% of the items on the test correct, we can infer that you know 85% of the subject of the test. This can only be done if the content domain (the subject of the test) is clearly defined and we have randomly sampled items from that domain for the test. Most high school assessment tests are criterion-referenced—in part because an individual’s score does not depend on how others do, but also because the content domain is defined in terms of a state curriculum and the items on the test are assumed to be a random sample of potential items from the curriculum.

Norm-referenced scores are those that indicate an individual’s relative position status in their peer population or compared to some other population—they tell us where the individual is in terms of the performance of others. The reference for the score is the norms. The interpretation directly depends on how others do on the test. The score simply tells us how an individual performed on the test relative to others who have taken the test.

*Norms*

An appropriate norm group must be recent, representative, and relevant.

Half of the students will necessarily be performing below the average (the norm).

*Relativity of Norms between tests*

1. test may differ in content
2. scale units may not be equivalent (SDs)
3. the standardization samples may not be equivalent

There are two types of norm-referenced scores: relative position status scores and developmental level scores.

# Relative Position Status Scores

Relative position status scores are also called *within-group norms*. These are norms of peers so that an individual’s performance is directly compared to the performance of others in the same peer group. There are several examples of these types of scores.

## Percentile scores

A percentile is the point on the distribution of scores below which a specific percentage of students score. If we were to say that an individual received a percentile score of 85, we could say that 85% of the population scored below the corresponding raw score.

## Standard scores

z-score, z = 

This is a summary indicator of the deviation (difference from the mean), which tells us the number of standard deviations a given score is from the mean. It is a standardized score (standardized by the standard deviation *SX*).

A z score of 0 is always the mean; a z score of +1.0 means that the score is one standard deviation above the mean. This transformation maintains the rank order of scores.

T scores, T = 10 *z* + 50

Some criticisms of z scores include the fact that there is a zero (indicating absence of something), negative numbers (with negative connotations), and decimal or fractional scores. To adjust z scores so that these problems are avoided, another standard score that could be used is the T score, which shifts the mean to 50 and the standard deviation to 10. This is common in psychological tests.

ETS scores = 100z + 500

These scores includes SAT, GRE, and other tests. They are simply standard scores scaled so that the mean is 500 and the standard deviation is 100 (similar to T scores but on a scale that is 10 times larger).

Deviation Quotient scores = 15z + 100

These scores are the typical scores used for IQ (intelligence) tests. Traditionally, IQ scores were computed by measuring the mental age of the subject and dividing that by their chronological age: mental age / chronological age x 100. This is not typically done these days. By employing standardized scores and rescaling, we obtain a score with a mean of 100 and a standard deviation of 15.

*Normal Curve Equivalents* (NCEs)

Normal curve equivalents were created by the federal government to standardize the reporting of Title I assessment and evaluation results from programs across the country employing various assessment instruments. These scores provide an equal-interval scale that allows statisticians to perform statistics on data that require interval level scales of measurement. These are similar to standardized scores presented earlier, but require normalized standard scores, which means that the scores are slightly transformed resulting in equal intervals rather than ordinal rank ordering of percentiles. The NCE scores correspond with percentiles at the 1st, 50th, and 99th percentiles.

NCE = 21.06 (normalized *z*) + 50

*Stanines* (standard nine)

Stanines were developed by the Air Force in WWII, based on the number of punches that could be made in one column of an IBM card – when data was entered by machine read punch cards. These scores are normalized with a mean of 5 and standard deviation of 2. They range from 1 to 9 with the following percent of the distribution within each of the nine stanines: 4%, 7%, 12%, 17%, 20%, 17%, 12%, 7%, 4%.

# Developmental Level Scores

Developmental level scores are also called *between-group norms*. These scores refer an individual’s performance to the performance of other groups, usually groups at a different age level or grade. There are several types of these scores.

## Grade Equivalent (GE)

Grade equivalent scores are based on the median (middle) performance of subjects in various grades. These scores are derived by administering a test to subjects in various grades and computing the median performance at each grade level—these provide the between-group norms. Then, when a subject takes the test, we look to see at what grade the median performance was the same as our subject—this is the GE. Interpretation is often problematic, because if a 3rd grade student obtains a GE of 4.8, it doesn’t mean that the student is performing at the level of a student in the 8th month of 4th grade. It actually means that the student is performing like a typical student in the 8th month of 4th grade on the 3rd grade test. Since the test is grade appropriate (3rd in our example), we cannot say that the student can successfully complete 4th grade work.

Problems include: (1) extrapolation beyond sample, (2) scores are not comparable across different grades regarding content mastered -- only within a grade, (3) it cannot be used as a standard because half of all sixth graders will be below the 6.0 GE, and (4) limited to primary education programs where subjects are common to a particular grade level.

## Mental Age

These are similar to GEs, except that various ages are used as norms rather than various grades. These scores are how the original Binet mental level scores were derived, which were used as the basis of IQ scores. These scores are not typically seen these days.

Give the test to subjects at different ages and plot the median raw score at each age group

## Scaled Scores

Many test publishers provide scale scores particular to their instrument. These often provide comparable results across forms at different levels for that particular assessment system. Each of these types of scores are described in the technical manual for each exam.

## Item-Response Theory (IRT)

Item response theory is a modern measurement theory that estimates ability in a way that is independent of the difficulty of the items that appear on the test. In a traditional classroom test, your score on the test is largely dependent on the difficulty of the items. If the items are more difficult, your score is lower. In IRT, ability is estimated in a way that is independent of the items on the test. These scores are now largely employed in all large-scale testing programs and provide the basis for computerized adaptive testing. The resulting scores are then scaled to meet the demands of the particular test (e.g., on a scale of 200 to 800 for ETS scores).

If the model fit the data, the results have the property of invariance: Item qualities are independent of the sample of participants used in calibration. Person scores are independent of the sample of items administered.

*Ordinal Scales*

Finally, there are also ordinal scales that are particular to the theory and the assessment instrument that employs them. One example includes the Piagetian developmental stages. These are simply ordered in terms of the hierarchical nature of Piagetian developmental theory, resulting in an ordering of developmental stages.

Criterion Referenced Scores

"Without a clearly defined domain of material to be tested, criterion referencing of the score is not possible."

### Summary of Standardized Scores: Positives & Negatives

|  |  |  |
| --- | --- | --- |
| **Scores** | **Positives** | **Negatives** |
| Percentiles | 1. Easy to compute 2. Easy to interpret 3. You can do intraindividual comparisons for performances of an individual on subscales and interindividual comparisons with others who took the same test | 1. Scores are simply ordinal, the indicate a rank-ordering of individuals, so the size of a percentile unit is not constant. A unit change in raw score is not constant across percentiles. 2. It is technically inappropriate to compute statistics employing percentiles, including the mean, standard deviation, and others. |
| Standard Scores  Deviation Quotients | 1. These result in the same distribution as raw scores. 2. When score distributions are normal (as they are with large norm groups), interpretation is easy, based on the normal curve. | 1. Interpretation is not straightforward with small samples. 2. In most cases, the distribution of scores changes as a function of the difficulty of specific forms, unless they have been equated. |
| NCEs | 1. Approximately equal interval, allowing for a full range of statistical analyses. | 1. Difficult to interpret, since only three points correspond to ease of interpretation: 1st, 50th, and 99th NCEs correspond to percentiles. |
| Stanines | 1. Quick classification of performances into 9 groups, where the 5th stanine is the median performance group. | 1. Not very precise since there is a wide range of scores that fall into each stanine. 2. Difficult to interpret since the percent of subjects in each stanine varies. |
| GEs | 1. Provide comparisons for performance between students at a given grade level and those in other grades. | 1. Misinterpretation is common, since many believe that the GE indicates the absolute level of performance. 2. Not useful to set standards or goals since, because of the way they are computed, only 50% can score above a grade equivalent: no more than 50% of 6th graders can perform above the 6th-grade equivalent. 3. GEs have little interpretative value beyond elementary school grades. They rely on common subjects within a grade level. |
| Scaled Scores | 1. Provide easy developmental comparisons, particularly if vertically scaled so that scores increase on the score scale across ages or grades. 2. Typically have equal units. | 1. Scales are often arbitrary and have no inherent interpretative value. 2. Typically have to read the technical manual to see how the particular scaled scores are derived and should be interpreted. |
| IRT Scores | 1. Equal units on the measurement scale, allowing the full range of statistical analyses. 2. Not dependent on the difficulty of the test or particular test form. 3. Can be used to create computerized adaptive tests which typically yield more precise estimates of ability. | 1. May be difficult to interpret because for some IRT models, the number correct does not correspond directly to the IRT score—it may be possible to obtain two different IRT scores from the same raw score if the items that were correct were different. |