



Definitions of Statistical Terms Relating to Tests



Log On. Let's Talk.

www.ets.org/letstalk

Listening. Learning. Leading.

Correlation

A correlation is a statistic that tells us how strongly two sets of measurements on the same individuals agree.

A correlation can be positive, zero, or negative. A positive correlation indicates that individuals who score relatively high on one measurement tend to score relatively high on the other. A negative correlation indicates that individuals who score relatively high on one measurement tend to score relatively low on the other. A zero correlation indicates that individuals who score relatively high on one measurement are about as likely to score relatively high as to score relatively low on the other.

Note that, where correlations are concerned, “high” and “low” are purely relative — relative to the average of the group. The entire group can be low on one measure and high on the other, but if the same individuals tend to be above the group average on both measures, the correlation will be positive.

The correlation can vary from -1.00 to $+1.00$. A correlation of $+1.00$ (or of -1.00) indicates that the relationship between the two measures is as strong as it can possibly be.

Do not make the mistake of thinking that whenever two measurements correlate highly, one can be used as a substitute for the other. Elementary school students’ reading test scores may correlate highly with their math test scores, but giving the students extra help and practice in math is not likely to improve their reading skills.

It is important to remember that, although a correlation tells how strongly two measurements tend to agree, it cannot tell why they agree. Two different characteristics may correlate highly, but the high correlation does not mean that one is a cause of the other. It often happens that students’ math scores correlate strongly with their writing scores. But that fact does not prove that skill in mathematics makes the students better writers, or that skill in writing makes them better at mathematics.

Criterion-referenced test

A criterion-referenced test is a test that is intended for comparing each test taker’s score with one or more fixed standards of performance. It is not designed primarily for determining each test taker’s relative position in a group. (A test that is intended for making those kinds of comparisons is called a

“norm-referenced” test.) For criterion-referenced testing, what matters is whether or not the test taker’s score meets the specified standard, regardless of how well or how poorly other test takers perform.

For any kind of testing, the primary concern in writing and selecting the questions is to measure appropriate content. However, criterion-referenced testing implies a different approach to test construction from norm-referenced testing. The purpose of the test is not to determine each test taker’s relative position in a group, but to determine whether each test taker is above or below a particular level of the knowledge or skill being measured. Therefore, the test makers try to write and select questions at the level of difficulty that will best separate the test takers with at least that level of knowledge or skill from those who fall short of that level. There is no need to determine which test takers are the strongest or which are the weakest. Therefore, the questions on a criterion-referenced test tend to have a narrower range of difficulty than those on a norm-referenced test.

Grade-equivalent scores

A grade-equivalent score expresses a student’s score on a test as being equal to the performance of a typical student at a particular grade-level. For example, a grade-equivalent score of 4.2 implies that the student has performed as well on the test as would a typical student in the second month of the fourth-grade year. (Conveniently, a school year has ten months.)

Grade-equivalent scores are based on the assumption that it is meaningful to define educational progress in terms of the grade-level at which an average student attains a given level of knowledge or skill. That assumption may be reasonable for subjects and grade-levels at which students make strong, steady progress from year to year, as they do in the lower elementary grades. But at higher grade-levels, there tends to be much less year-to-year improvement in the average student’s knowledge or skill. An average second-grader reads much better than an average first-grader, but an average eighth-grader reads only slightly better than an average seventh-grader. Therefore, a difference between grade-equivalent scores of 1.5 and 2.5 in reading is an important difference, while a difference between grade-equivalent scores of 8.5 and 7.5 in reading is not. In interpreting grade-equivalent scores, it is important to remember that a difference of one unit means more at the lower grades than at the higher grades.

One of the greatest problems with grade-equivalent scores is that they are easily misinterpreted. A grade-equivalent score of 4.0 does not represent the level of performance to be expected from most beginning fourth-graders. It represents the level of performance of an average beginning fourth-grader. Half (or nearly half) of all beginning fourth-graders will have grade-equivalent scores below 4.0, because that is the way the grade-equivalent score scale is defined. But many people are unwilling to accept the notion of half the students in the country performing “below grade level.”

Another common misinterpretation is that the grade-equivalent score represents the level of work a student is ready to do. But, for example, if a beginning third-grade student earns a grade-equivalent score of 5.0 in math, that score does not mean that the student is ready for fifth-grade work in math. It means only that the student did as well on the math test for beginning third-graders as would a typical beginning fifth-grader. The test for beginning third-graders will not include many questions on skills that are taught in the fourth-grade, but a student who has not yet learned those skills is not ready for fifth-grade work.

Mean

The mean score is the average score of a group of test takers, computed in the usual way — by adding up all the scores and dividing by the number of test takers. The mean score is the most common statistic for describing the performance of a group of test takers with a single number.

One reason for using the mean score to describe the performance of a group is that every individual’s score has the same effect on the group mean score. This feature of the mean score makes it especially useful for before-and-after comparisons. However, this same feature can make the mean score a poor statistic for describing the performance of a typical test taker in the group. If most of the test takers in the group have high scores, just a few test takers with very low scores can have a substantial effect on the group mean score. Other statistics, such as the median will not be affected by a few test takers that are very different from the rest.

Measurement error

Measurement error is the term that people in the testing profession use to refer to factors that can

make test scores less than perfectly reliable. Unfortunately, the term “measurement error” can be misleading. To say that a test score contains measurement error does not mean that someone has made a mistake in administering or scoring the test. It means only that a test taker’s score is influenced by factors that affect test takers at random.

An individual test taker’s score will depend on which specific questions the test taker was actually asked. The same test taker, called on to answer a different, equally difficult set of questions measuring the same types of knowledge, might perform somewhat differently. From the test taker’s point of view, the selection of the particular questions on the test is determined by the luck of the draw. Therefore, the selection of specific questions that appear on the test is referred to as a source of measurement error.

Similarly, if the scoring of the test requires human judgment, an individual test taker’s score will depend on the specific scorers who evaluated the test taker’s responses. From the test taker’s point of view, the selection of those particular scorers is determined by the luck of the draw. Therefore, the selection of scorers is also a source of measurement error.

The selection of a particular day for testing is also a source of measurement error, because a test taker might perform differently on a different day, even if no real learning or forgetting occurred.

Every reliability statistic refers to one or more sources of measurement error. The reliability statistics computed for most large-scale tests refer to a single source of measurement error: the selection of specific questions. If the scoring of the test involves judgment, the reliability statistics should also include a second source of measurement error: the selection of scorers.

The statistics that describe the extent to which the test takers’ scores are influenced by a particular source (or combination of sources) of measurement error are called the “standard error of measurement” and the “reliability coefficient.”

Median

The median is a statistic that describes the performance of a typical test taker in a group of test takers. The median is the test score that separates the upper half of the group from the lower half. It is also referred to as the “50th percentile.”

The median is not affected by changes in the

performance of the strongest students. It is not affected by changes in the performance of the weakest students. It is affected only by changes in the performance of the students whose scores place them in the middle of the group. This characteristic of the median makes it a very good statistic for describing the performance of a typical test taker, but not always a good statistic for making before-and-after comparisons. For before-and-after comparisons, it is usually better to use a statistic that is affected by all the test takers' performance, such as the mean score.

Norms

One way to make test scores meaningful is to provide information about the scores of a group of test takers — not just any group, but a group that people might want to know about. This group is called a “norm group.” The norm group could consist of the students in a particular grade in a particular school district, or in an entire state, or in the entire nation. Or it could be simply all the test takers who happen to take the test. Norms are statistics that describe the performance of the norm group. The norm group can include students who did not actually take the test. In that case, the statistics must be estimated by giving the test to a sample of the test takers in the norm group.

If the norm group consists of the test takers at those institutions that use the test, the norms are referred to as “user norms.” If the norm group is limited to a single school district, the norms are referred to as “local norms.” If the norm group includes test takers from an entire state, the norms are referred to as “state norms.” If the norm group includes test takers from the entire nation, the norms are referred to as “national norms.” In interpreting norms, it is important to know what norm group they refer to. User norms can be very different from national norms. Local norms can differ substantially from state or national norms.

The norms — the statistics that describe the norm group's performance on the test — can be presented in a number of different ways. Possibly the most common way is to show, for each possible score on the test, the percentile rank of that score in the norm group — or in two or more different norm groups. A test score can have one percentile rank in a local norm group and a very different percentile rank in a national norm group. Another common way to present norms data is to show selected percentiles of the test scores of one or more norm groups.

Norm-referenced test

One way to make test scores meaningful is to provide information about the scores of a group of test takers. Often, the test scores are reported in ways that are intended to make these kinds of comparisons easy to do. The group with which test users are encouraged to compare each test taker's score is called a “norm group.” Test scores that are intended to be interpreted in this way are called “norm-referenced scores.” And a test that is intended primarily for comparing each test taker with a group of test takers is called a “norm-referenced test.”

For any kind of testing, the primary concern in writing and selecting the questions is to measure appropriate content. However, norm-referenced testing implies a particular approach toward the writing and selection of questions for the test. Determining a test taker's relative position in a group will be easier to do if the scores of the group are spread widely over the range of possible scores. One way to spread out the test takers' scores is to make the test takers answer lots of questions. However, the number of questions that can be included on a test is limited by the testing time available. A question that all test takers answer correctly does not help to separate the test takers who know more from those who know less. Neither does a question on which all the test takers perform as if they were guessing at random. Therefore, the makers of norm-referenced tests concentrate on “middle-difficulty” questions — questions that many test takers will answer correctly, but many others will not. For almost any type of knowledge or skill, it is possible to write both easy questions and hard questions. If even the weakest test takers can answer a question correctly, the test makers rewrite it to make it harder. If they cannot make the question harder, they replace it with a harder question testing a different point of knowledge in the same category or a different application of the same skill. Similarly, questions that even the strongest test takers cannot answer correctly are revised or replaced with easier questions.

For many years, the norm-referenced approach to testing was the only approach that was taken seriously by people in the testing profession. However, there is now an alternative approach, called “criterion-referenced testing.”

Percentile

A percentile is a number that gives one specific piece of information about the scores of a group of test takers. A percentile is the score that separates a specified percentage of the test takers — those with lower scores — from the rest of those with higher scores. For example, the 30th percentile is the score that separates the lowest-scoring 30 percent of the test takers from the highest-scoring 70 percent. The 90th percentile is the score that separates the lowest-scoring 90 percent from the highest-scoring 10 percent.

Percentiles are a particularly useful way to describe the scores of a group, because they can show how the strongest students, the above-average students, the average students, the below-average students, and the weakest students in the group performed.

Some percentiles have special names. The 50th percentile, which separates the lower half of the test takers from the upper half, is called the “median.” The 25th percentile, which separates the bottom quarter of the test takers from the rest, is called the “first quartile.” Similarly, the 75th percentile is called the “third quartile.”

The concept of a percentile is closely related to that of a percentile rank. The 40th percentile of a group is the score that has a percentile rank of 40 in that group; the 68th percentile is the score that has a percentile rank of 68; and so on.

Percentile rank

The percentile rank compares a test taker’s performance on a test with the performance of a group of people who took that test. The test taker’s percentile rank is the percentage of the group who earned lower scores than the test taker. (Often, it also includes half the percentage that received exactly the same score as the test taker.)

A test taker’s percentile rank depends on the group. In a stronger group, there will be fewer low scores, and the test taker’s percentile rank will be lower. In a weaker group, there will be more low scores, and the test taker’s percentile rank will be higher. A test taker’s score could have a percentile rank of 35 in the test taker’s own school, a percentile rank of 52 in the school district, and a percentile rank of 46 in the entire state. The test taker’s percentile rank is meaningful only if you know what group of test takers it refers to.

A percentile rank is not the same as the percent of test questions that the test taker answered correctly. The percentile rank refers to a percentage of people — the test takers in a group. If the test is easy for the group, a test taker can answer 70 percent of the questions correctly and get a score at the 10th percentile or even lower. If the test is hard for the group, a test taker can answer 70 percent of the questions correctly and get a score at the 90th percentile.

Looking at differences between percentile ranks can be misleading. The reason is that test takers’ scores tend to be bunched closely together in some parts of the score range (usually the middle) and spread out in other parts of the score range (the high and low ends, as in the familiar “bell curve”). Where test takers’ scores are bunched closely together, one or two additional correct answers can move the test taker ahead of a substantial number of people. As a result, the test taker’s percentile rank will change substantially. But in the portions of the score range where there are not many scores, one or two additional correct answers will not move the test taker ahead of many people, and the test taker’s percentile rank will change only slightly.

The concept of a percentile rank is closely related to that of a percentile. The 40th percentile of a group is the score that has a percentile rank of 40 in that group; the 68th percentile is the score that has a percentile rank of 68; and so on.

Reliability coefficient

The reliability coefficient is a statistic used to describe the reliability of the test scores of a group of test takers. The reliability coefficient is the correlation of the scores the test takers would receive if they were tested twice — with different specific questions, with different scorers scoring the responses, etc. The reliability coefficient can vary from .00 to 1.00. It describes the extent to which the test takers, whose scores were above average on the first testing, would tend to score above average, to the same degree, on the second testing (and similarly for the test takers whose scores were below average).

A reliability coefficient always refers to a specific test, a specific group of test takers, and a specific source (or combination of sources) of measurement error. The reliability coefficient measures the consistency with which the test distinguishes between the stronger test takers and the weaker test takers in

the group. The reliability coefficient can vary greatly from one group of test takers to another. If the test takers in the group differ widely in the subject tested, it will be easy to distinguish the stronger test takers from the weaker ones, and the reliability coefficient will tend to be high. If the test takers in the group do not differ greatly, it will be much harder to distinguish the stronger test takers from the weaker ones, and the reliability coefficient will tend to be low.

Scaled scores

A scaled score is a score that is expressed by a set of numbers chosen (often arbitrarily) by the testing organization. A scaled score cannot be determined simply by counting the test taker's number of correct answers or by summing the ratings awarded to the test taker's responses. Those kinds of scores — number correct, sum of ratings, percent of possible points — are called “raw scores.” Raw scores depend heavily on the difficulty of the questions or problems on the test. But in large-scale testing programs, the questions on the test change from one testing session to another. To report raw scores would not be fair to a test taker who happened to get a difficult set of questions. Instead, most testing organizations report scores that are statistically adjusted to compensate for changes in the difficulty of the questions. Typically, they report the adjusted scores by using a range of numbers that is different from that of the raw scores. This range of numbers is called the “score scale,” and the scores are called “scaled scores” (or, sometimes, “scale scores”). Scaled scores make it possible to compare the scores of test takers taking different editions of the test. These kinds of comparisons are important for making some kinds of decisions (e.g., college admissions) and for measuring the progress of groups of students.

Different testing programs use different score scales for their tests. Here are some examples:

ACT scale: 1, 2, 3, 4, 5, ... , 34, 35, 36

Praxis scale: 100, 101, 102, ... , 198, 199, 200

PPST scale: 150, 151, 152, ... , 188, 189, 190

SAT scale: 200, 210, 220, 230, 240, ... , 780, 790, 800

In interpreting a test score, it is important to know the score scale for the test. A score of 200 is the lowest score possible on the SAT scale but the highest score possible on the Praxis scale. A score of

155 would be near the middle of the Praxis scale but near the bottom of the PPST scale.

Sometimes a testing organization selects a particular score scale to make sure that the scaled scores will not be mistaken for other types of scores, such as the number correct, or percent correct, or percentile rank .

To determine a test taker's scaled score on a test, the first step is to determine the test taker's raw score. Then the raw score is translated into a scaled score, according to a table or a formula called the “raw-to-scale conversion.” Because the questions on the test change from one edition of the test to another, there is a separate raw-to-scale conversion for each edition of the test. If the questions on one edition of the test are harder than those on another edition of the test, that difference in difficulty will be reflected in the raw-to-scale conversions. To earn a particular scaled score, a test taker who takes the edition with easier questions will have to answer more questions correctly than a test taker who takes the edition with harder questions. The raw-to-scale conversions that make scores comparable across different editions of the test are determined by a statistical process called “score equating.”

Standard deviation

The standard deviation is a statistic that tells how much the scores of a group of test takers differ from each other. If the scores are widely spread out, the standard deviation will be large. If the scores are all bunched closely together, the standard deviation will be small. If all test takers have exactly the same score, the standard deviation will be zero.

The standard deviation is expressed in the same units as the test scores themselves. It can be interpreted as an answer to the question, “On the average, how much do the individual scores differ from the average score for the group?” (However, the calculations are a bit more complicated than this interpretation would imply.)

Sometimes the standard deviation is used as a way of expressing differences between scores. For example, suppose the standard deviation of the scores of a group of test takers is 10 points. If two of the test takers have scores that differ by five points, they might be described as differing by “half a standard deviation.” Because the standard deviation often differs from one group to another, it is important to identify the group when comparing scores in this way.

The standard deviation is also used, along with the group mean score, to describe a test taker's relative position in a group. If a test taker's score is 6 points above the group mean score, that test taker's score would be described as "0.6 standard deviations above the mean." Again, it is important to identify the group when describing scores in this way.

One reason for describing and comparing scores in terms of the standard deviation is that, for a person who is not familiar with scores on a particular test, "half a standard deviation" is more meaningful than "five points." Another reason is to compare scores on two tests that use completely different score scales. A third reason is to focus on the test taker's relative position in a particular group.

Standard error of measurement

The term "standard error of measurement" does not mean that somebody has made a mistake in administering or scoring a test. The standard error of measurement is a measure of the extent to which test takers' scores would vary over many repeated testings, with different questions, scorers, etc. These factors, which make test scores less than perfectly reliable, are referred to by people in the testing profession as sources of measurement error. The standard error of measurement always refers to a particular source (or combination of sources) of measurement error. If that source of measurement error has only a small effect on the test scores, the standard error of measurement will be small, and the scores will be highly reliable. If that source of measurement error has a large effect on the test scores, the standard error of measurement will be large, and the scores will not be highly reliable. Therefore, the standard error of measurement is a useful way to measure the reliability of the scores. The standard error of measurement is expressed in the score scale of the test.

In any group of test takers, the standard error of measurement of scores on a test is closely related to the reliability coefficient. However, unlike the reliability coefficient, the standard error of measurement tends to vary only slightly from one group of test takers to another.

Stanine

A stanine is a type of score that compares a test taker's performance on a test with the performance of a group (called the "norm group"). The stanine places the test taker into one of nine levels of the

norm group. Stanine 1 consists of the test takers in the norm group with the lowest scores; Stanine 2 consists of the test takers with the next lowest scores; and so on, up to Stanine 9. The following table shows the approximate percent of the norm group in each stanine:

Stanine	Percent of norm group
9	4
8	7
7	12
6	17
5	20
4	17
3	12
2	7
1	4

Notice that most of the test takers are in the middle stanines — Stanines 4, 5, and 6. There are very few test takers in Stanines 1 and 9. Typically, when a large group of test takers takes a test, there are many test takers with average and near-average scores and few test takers with very high or very low scores (giving rise to the familiar "bell-shaped curve"). However, even if the actual test scores of the norm group do not form the usual bell-shaped pattern, their stanine scores will. The stanine scores of the norm group must show this pattern, because the borders of the score categories are chosen to put a fixed percentage of the norm group into each stanine.

Although the percentage of the norm group in each stanine is set in advance, the percentage in each stanine for some other group of test takers can be quite different from that of the norm group. A stronger group will have more test takers in the higher stanines and fewer in the lower stanines. A weaker group will have more test takers in the lower stanines and fewer in the higher stanines.

The stanine is not a very precise measure. Two test takers can have scores that are very close together but in different stanines, if one score is just below the boundary between the two stanines and the other score is just above the boundary. Two other

