Pearson

Standardized Clinical Assessment for Practitioners: A Primer

Authored by Pearson Clinical Assessment's Scientific Council

2023

psy·cho·met·rics

/ˌsīkəˈmetriks/

noun

the science of measuring mental capacities and processes.

Psychometrics is the study of the measurement of human behavior, concerned with constructing reliable and valid instruments, as well as standardized procedures for measurement.

What distinguishes a standardized assessment from a nonstandardized assessment?

Standardized test

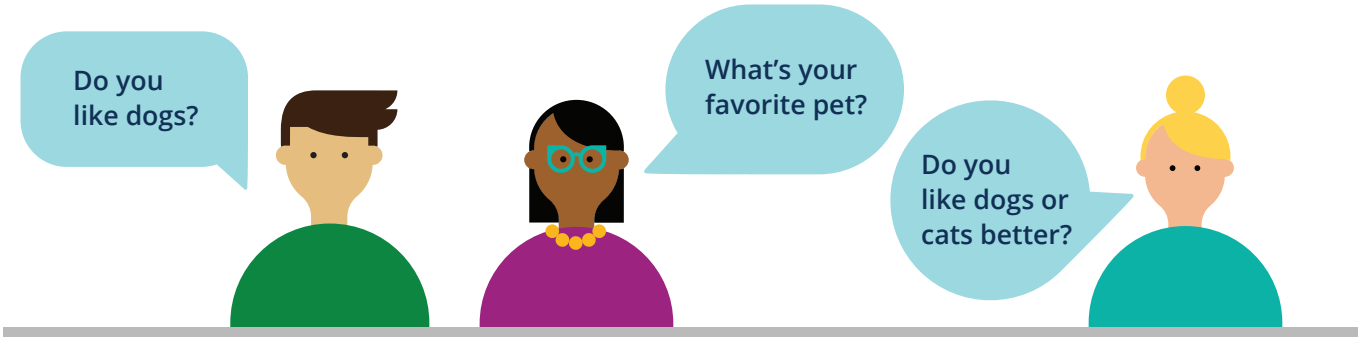
Standardized tests require strict adherence to administering and scoring procedures by every professional who uses the test. Additional characteristics include:

- Data have been collected on large numbers of participants
- Data determine the average score (mean) and the standard deviation, which the clinician then uses to benchmark the performance of the client tested
- Structured procedures for interpreting results usually involve comparing a client's score to the scores of a representative sample of people with similar characteristics (e.g., same-age peers)

Nonstandardized test

Nonstandardized tests are usually created to measure constructs that are clinically meaningful when no standardized test is available. However, test use and interpretation can be problematic when:

- The test is not developed based on data from a large number of subjects who were tested and whose performance was scored in exactly the same way by each clinician, according to a structured set of rules
- There is no scientifically supported basis for what constitutes a good or poor score, and other clinicians administer the same test, but score it based on their own criteria
- No data have been collected to verify the appropriateness of the administration procedures, the test items, and the scoring



Why is standardized assessment important?

Have you ever been in a situation in which two people have asked a very similar question of the same person, but they each received different responses? Perhaps the two people were asking *essentially* the same question in slightly different ways or in different contexts. The exact wording of a question or the order in which questions are asked can influence the response. In a testing situation, the relationship between the examiner and examinee can influence the level of effort the examinee puts into the test. To administer standardized assessments, examiners are trained to ask each client identical questions in a specific order and with a neutral tone to avoid inadvertently influencing the response. Examiners are trained to score responses in a uniform way so that one examiner does not rate a particular response as within normal limits while another examiner rates the very same response as outside of normal limits. This is important because many tests elicit verbal responses from examinees and have to be judged for accuracy. Standardized scoring rules give examiners a common set of rules so they can judge and score responses the same way.

7 reasons why standardized tests are an important part of clinical assessment practices

1. Help you gather and interpret data in a standard way
2. Provide evidence to support or disconfirm your hypotheses
3. Support requests for services or reimbursement
4. Identify patterns of strengths and weaknesses to help guide the development of an appropriate intervention and treatment plan
5. Measure treatment outcomes
6. Provide a consistent means to document client progress
7. Gather a body of evidence that can be disseminated as a set of best practice recommendations

How are test scores useful for outcomes-based practice?

Outcomes measurement can inform and improve your practice. Knowing the location of a person's score on the normal curve enables you to determine their unique starting point prior to therapy. Following a course of treatment, the client can be retested. If the starting point was below average, and the retest score is in the average range, then there is clear documentation that the client improved relative to the norm group. However, if the second score is within the standard error of measurement of the first score, then there is no clear evidence of treatment effectiveness relative to the norm group. In this situation, the examinee might have made progress in their skills; however, the examinees in the normative sample also showed progress over time. As a result, the difference between the examinee's standard scores over time was minimal. In contrast, growth scale values (GSVs), which are available on many standardized tests, indicate whether a client's performance significantly improved or declined relative to their own past performance.

Assuming the length of treatment was adequate and applied properly, test scores such as standard scores and GSVs can help the outcomes-based therapist to decide whether to continue the current treatment or consider an alternative.

Types of scores in standardized testing

Raw score.

The subtest raw score is the sum of the item scores. Raw scores are not directly interpretable, and they are not comparable from one subtest to the next. Raw scores do not even communicate well within a particular subtest because the same score may be high or low depending on the examinee's age/grade. Raw scores are typically converted to a standard score that has uniform meaning across subtests and ages/grades.

Standard score.

Standard scores provide a common metric that reflects how the examinee's performance compares with that of grade- or age-matched peers. "Standard scores" are standard because each raw score has been transformed according to its position in the normal curve so that the mean and the standard deviation *(SD)* are predetermined values (e.g., mean of 100 and *SD* of 15). Standard scores enable test performance to be interpreted based on a normal distribution (normal curve). Unlike raw scores, standard scores are on an equal interval scale, so the size of difference between two scores represents the same amount of difference in the skill being measured regardless of where on the scale the scores fall.

Percentile ranks.

Percentile ranks are also commonly used to interpret test results, and they link directly to the standard scores based on the normal curve. A percentile rank indicates the percentage of people who obtained that score or a lower one. So, a percentile rank of 30 indicates that 30% of individuals in the standardization sample obtained that score or lower. Similarly, a percentile rank of 30 indicates that 70% of individuals in the standardization sample scored higher than that score. Percentile ranks range from 1 to 99, with 50 as the median. Unlike standard scores, percentile ranks do not have equal intervals.

Growth scale values.

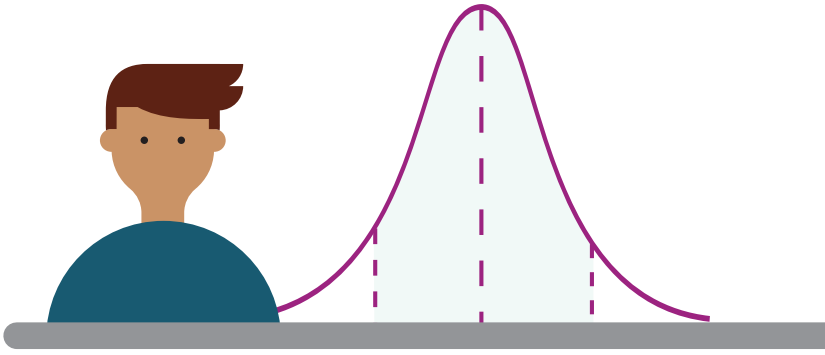
Growth scale values (GSVs) are preferred for measuring growth because GSVs reflect the examinee's absolute (rather than relative) level of performance. GSVs are useful for comparing a client's performance compared to their own past performance. Unlike standard scores and percentile ranks, GSVs do not compare a client's performance to that of a reference group. Raw scores by themselves are undesirable for measuring growth because they are not on equal interval scales, which means that a given difference in raw score points does not have the same meaning at different score levels.

Why must I convert raw scores into another score?

Raw scores need to be transformed into standard scores to compare a client's performance to the performances of other examinees. For example, say you have tested a second-grade boy named Jamal and administered all the test items precisely according to the test directions, in prescribed item order, and have followed the start/stop rules and scoring directions exactly. Jamal receives a raw score of 32. How do you know if this score is high, low, or average? First, you would want to know the average score for second-graders (children Jamal's age). If the average (or mean) score for second graders is 40 points, you know that Jamal's score is lower than average, but you still need to ask, "Is it very low or just a little bit low?" To answer this question, psychometricians use the test's *standard deviation.* The standard deviation is derived from the test's normative data, using a complex statistical formula. It basically tells us how much variability there is across the scores of the subjects tested in the normative sample.

Let's say the standard deviation of this test is 4 points. A raw score of 36 would be one standard deviation below the mean of 40. With this information, we know that Jamal's score of 32, which is *two* standard deviations below the mean, is very low. If the standard deviation of the test was 10 points, then we would say that Jamal's score of 32 is less than one standard deviation below the mean—which is not very low.

Psychometricians develop norms to convert each raw score (i.e., number of items correct) into a standard score. All of these statistical adjustments have already been taken into account by the subtest norms.

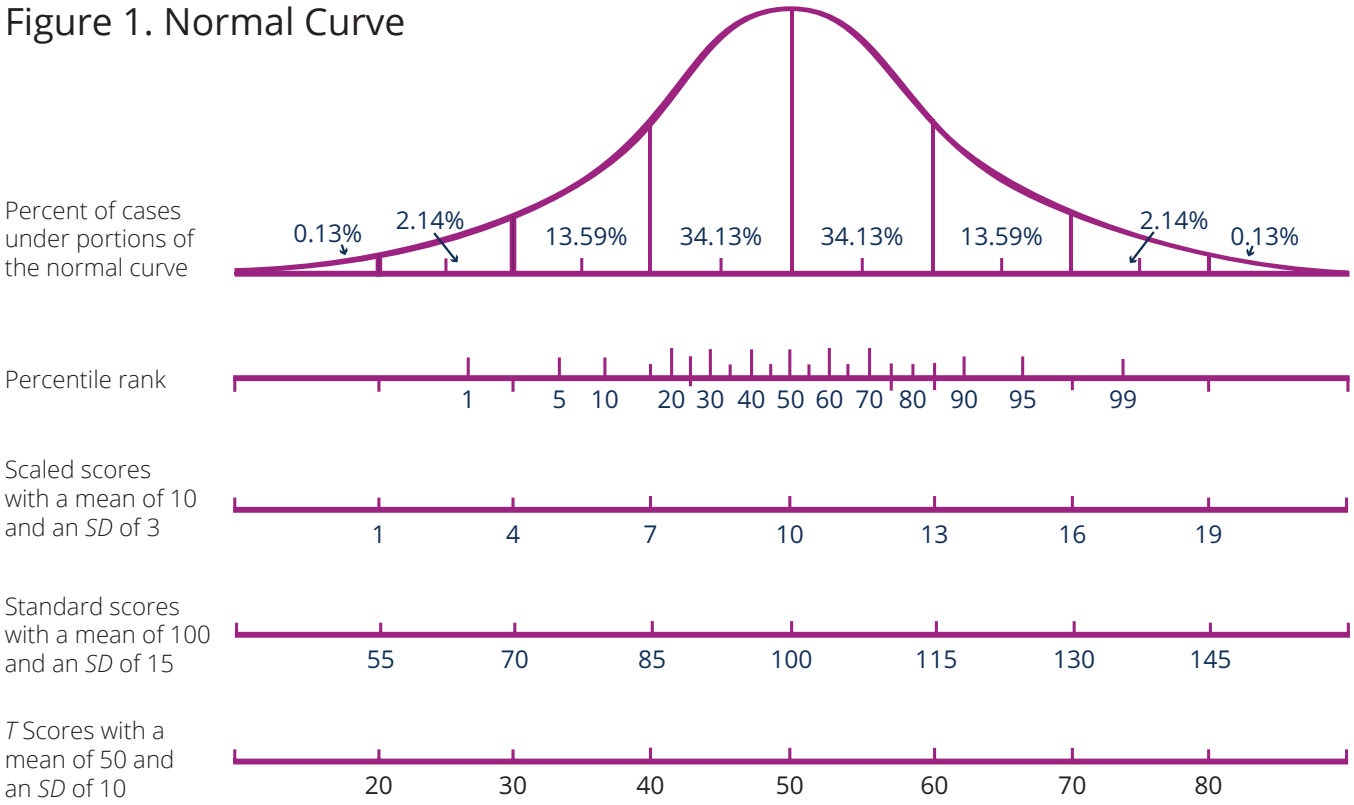


How do standard scores relate to the normal curve?

Standard scores are "standard" because the original distribution of raw scores have been transformed to produce a normal curve (a standard distribution having a specific mean and standard deviation). Figure 1 shows the normal curve and its relationship to percentile ranks and several types of standard scores, including scaled scores with a mean of 10, standard scores with a mean of 100, and T scores with a mean of 50. As shown, the mean is the 50th percentile. This means that 50% of the normative sample obtained this score or lower. One and two standard deviations above the mean are the 84th and 98th percentiles, respectively. One and two standard deviations below the mean are the 16th and 2nd percentiles. While one standard deviation below the mean may not sound very low, it actually means that this client's score is better than only 16% of their peers.

Standard scores are "standard" because the normative data have been transformed to produce a normal curve

Figure 1. Normal Curve



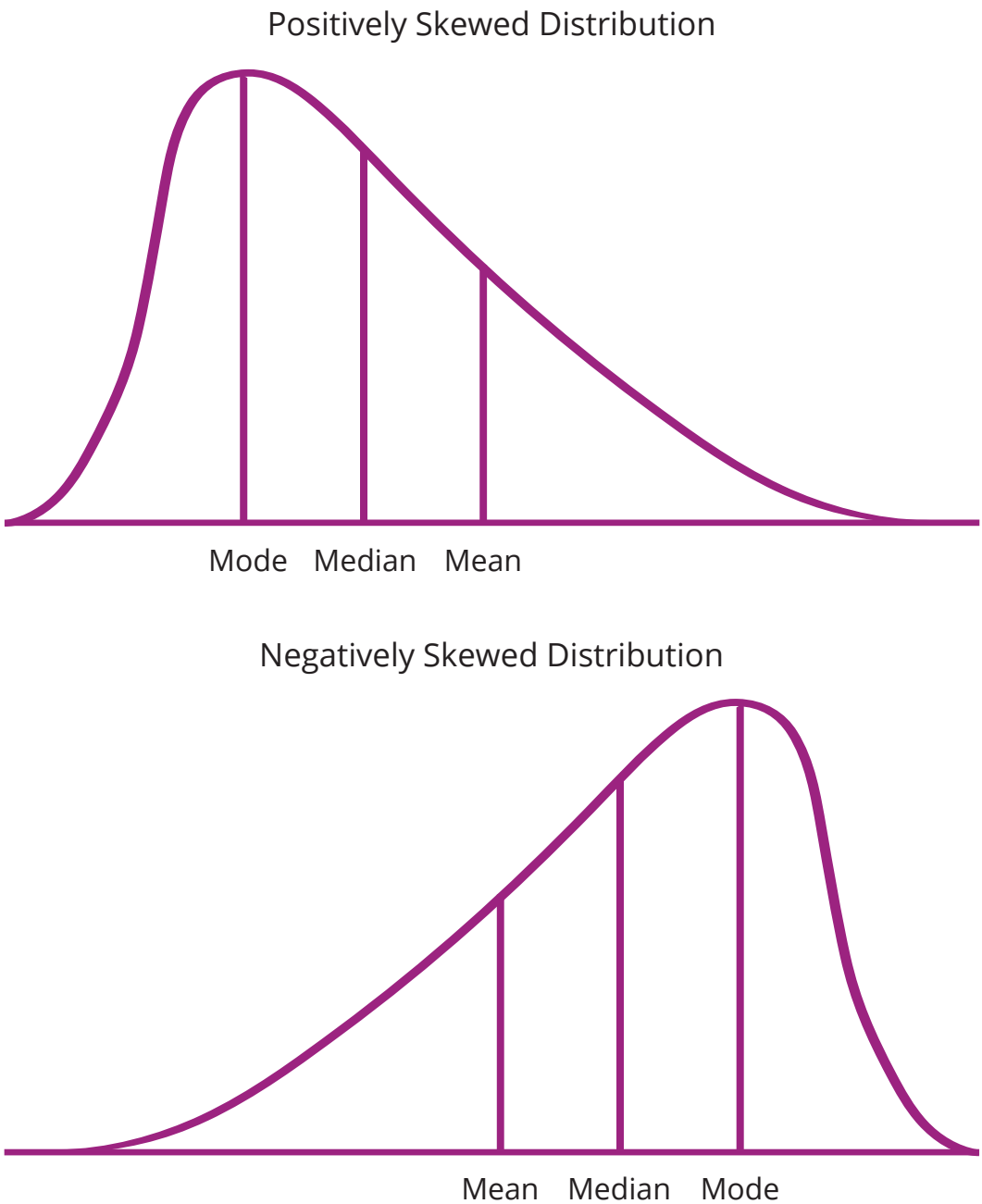
Standard scores are assigned to all raw scores based on the standard deviation, but there are many different types of standard scores. One common type of standard score is a metric where the mean is 100 and the standard deviation is 15. In the example of Jamal, the raw score mean was 40 points and the raw score standard deviation was 4 points. Jamal's obtained raw score of 32 is two standard deviations below the mean at the 2nd percentile, so it would be assigned a standard score of 70. A standard score of 70 means the same thing for all tests normed using this 100/15 metric. In this way, standardized tests make it easier for you to interpret scores on different tests and to compare scores across tests.

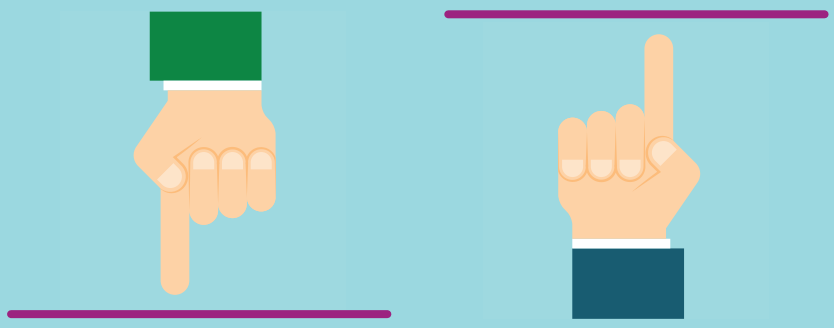
Another popular standard score metric is the T score. In this system the mean is always set to *T* 50, and the standard deviation is always 10 *T* score points. So, *T* 40 and *T* 60 are one standard deviation below and above the mean, respectively. If Jamal's raw score had been transformed to a *T* score metric, it would be *T* 30, which has the same meaning as a standard score of 70 (i.e., two standard deviations below the mean at the 2nd percentile).

For normally distributed constructs, percentiles and standard deviations line up as shown in Figure 1 (i.e., one standard deviation below the mean is the 16th percentile). However, keep in mind that not all constructs of clinical interest are normally distributed in the population. When a construct is distributed in a way that the scores pile up on one end of the scale and taper off gradually at the other end, the distribution is called *skewed.* These distributions can be either positively or negatively skewed, as shown in Figure 2. A negatively skewed distribution might be obtained when measuring a construct that most subjects of one age can easily perform and only very few cannot. For example, a test of rhyming for 8-year-olds might be negatively skewed because most 8-year-olds can easily perform these tasks, and only very few cannot. A positively skewed distribution may be obtained when measuring a construct that most individuals cannot perform and only a few can. For example, a test of phonological awareness for 3-year-olds may be positively skewed because most cannot perform these tasks, but a few can. In skewed distributions, the percentiles and standard deviation units do not line up as they do in a normal curve. They vary to the extent that the distribution is skewed.

For example, a test of rhyming for 8-year-olds might be negatively skewed because most 8-year-olds can easily perform these tasks, and only very few cannot.

Figure 2. Positively and Negatively Skewed Distributions





Why do many tests have basal and ceiling rules?

Many tests are designed for assessing clients across a wide range of ages and abilities; therefore, not all test items are necessary or appropriate for every client. In most cases, the test items are ordered from easiest to hardest.

Basal rules enable you to establish where to start the test so that you do not need to administer every item. For example, if you are testing a 6-year-old child for language development, the test developers might have you start the test with items appropriate for 5 ½-year-olds just to be sure the child understands the task and to provide the child some practice with the items. You would not need to administer items intended for 3- or 4-year-olds unless the child had trouble responding to items designed for 5-year-olds. Typically, the *start point* in any test, subtest, or series of items is set at a level where 90% or more of all examinees that age have responded to the earlier items correctly. This helps reduce testing time and ensures that you administer only the items appropriate for each examinee's ability level.

Ceiling rules enable you to know when to stop testing because you have exceeded the examinee's ability to respond correctly. Psychometricians have analyzed the standardization data to determine when you can be sure that if you administer another item the examinee will likely get it wrong. Usually, the *discontinue rule* is set so that after a certain number of items are answered incorrectly there is less than a 10% chance that the examinee will respond to any of the remaining items correctly. This reduces testing time, but equally importantly, it prevents frustrating the examinee by administering items that are too difficult for them.

What are confidence intervals, and why should I use them?

Because we are measuring human behavior and not a physical characteristic (e.g., height or weight), there is always some measurement error inherent in all clinical tests. Sources of measurement error include fluctuations in human performance over time related to health or fatigue, lack of internal consistency within a set of questions, or even differences in rapport between the examiner and examinee.

For all these reasons, the examinee's true score may be slightly higher or lower than the specific score obtained. It is best to think of a range of scores that most likely describe the examinee's performance, rather than a single point score. The *confidence interval* is a range of scores around an obtained score that is sure to include the examinee's true score with 90% or 95% likelihood. Many tests report *critical values* that may be used to establish a confidence interval, such as plus or minus 5 points. Confidence intervals are derived from the *standard error of measurement.*

What is standard error of measurement and why should I be concerned about it?

The **standard error of measurement *(SEM)*** is an estimate of the amount of measurement error in a test, which is different for every test. Conceptually, the *SEM* is the reverse of reliability—the greater the reliability of a test, the smaller the standard error of measurement.

You should be concerned about standard errors of measurement because you can have more confidence in the accuracy of a test score when the reliability is high and the standard error of measurement is small. Psychometricians use the *SEM* to create the confidence interval. The higher the reliability, the smaller the *SEM* and the narrower the confidence interval. A narrower confidence interval means you have a more precise score. We recommend that practitioners take measurement error into account when interpreting test scores by using confidence intervals. Some tests have confidence intervals built into the norms tables.

How do I determine if a test has "good" norms?

Size of normative sample

The accuracy of any standard score depends on the accuracy of the raw score mean and standard deviation obtained from the normative sample used to create the transformations to standard scores. The normative sample must be large enough to provide stable estimates of the population mean score and standard deviation. Very small normative samples may not have accurate raw score means and standard deviations because too much depends on the performance of the few subjects tested. The larger the sample, the more confidence you can have that a few errant subjects (referred to as *outliers*) did not have undue influence on the raw score mean and standard deviation. We can then say that the raw score means and standard deviations obtained from the normative data are stable.

Sample representation

There is more to quality norms than the size of the sample. The subjects in the sample must be representative of the types of clients with whom you use the test. Representation, however, is sometimes misunderstood. A test does not have to include examinees with a particular clinical condition or disorder in the normative sample for that test to be used fairly and appropriately in an evaluation for that disorder.

Factors known from previous research to affect performance on the task of interest should be represented in the normative sample. For example, it is known that demographic and cultural differences can impact a child's cognitive and language development. When creating a test of early development, it is important to ensure that children from different backgrounds are represented in approximately the same proportions as they are found in the general population. It is incumbent upon the test developer to understand what factors influence scores on the construct being measured and ensure proper representation of those factors in the normative sample.

Age of normative sample

Norms that were collected many years ago may no longer fairly represent today's population of children or adults. In the area of cognitive assessment, researchers have shown that norms tend to shift approximately 3 to 4 points every 10 years. Cognitive ability test scores typically improve across generations due to societal improvements in neonatal care, well-baby checks, nutrition, education, etc., so the norms from 10 years ago may no longer apply. As a general rule, more recent norms are preferred because they are more reflective of the current population. For this reason, test publishers may periodically update the norms of standardized tests.

Do all assessments require norms?

Not all tests require norms. When tests are keyed to specific external standards or criteria they are called *criterion-referenced tests.* This is common in educational settings where students must meet curriculum standards set by the state board of education. In clinical practice, some constructs may be better referenced to an external standard of expected performance than to a sample of normal subjects. For example, first grade students may have a goal to read a certain number of words correctly from a graded word list.

Types of reliability

In general, *reliability* refers to the dependability of a test over time. Actually, there are several different types of reliability and each type estimates a different source of possible measurement error. The measures of reliability all range between 0 and .99. Two types of reliability that are commonly reported include the following.

Internal consistency reliability

This measures the extent to which all the items in a test measure the same construct. To calculate internal consistency reliability, psychometricians use various formulas such as split-half reliability, or the *coefficient alpha* (also called Cronbach's Alpha). All of these formulas are based on some way of calculating the extent to which the items in a test correlate with each other. The higher the correlation between items, the more we can assume that all the items measure the same thing. So, this type of reliability estimates measurement error based on inconsistency within the item set. For test batteries that include multiple subtests, this should be calculated separately for each subtest.

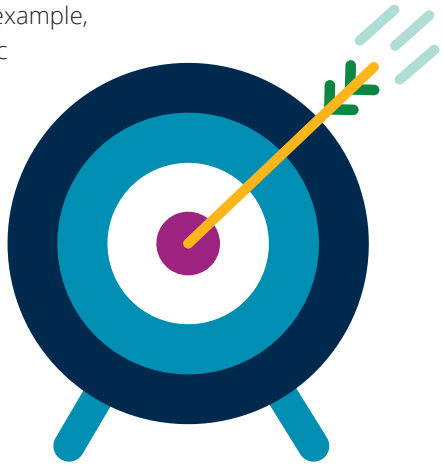
Test–retest reliability

To estimate this type of reliability, the same test is administered twice to the same examinee, with a specific interval between the two administrations. Scores from the two test administrations are compared to see how highly they correlate and how much change there is between the scores in the two testing sessions. This type of reliability estimates measurement error from changes in human performance over time and is sometimes referred to as the *stability coefficient.*

What you need to know about validity

Though internal consistency reliability is a way to determine if all of the items in a test measure the *same* thing, other information is collected to provide evidence that the items measure the *right* thing. In other words, does a test of verbal intelligence actually measure verbal intelligence or is it really measuring language proficiency? To answer this question, one might design a study to show that a new verbal intelligence test correlates highly with other established tests of verbal intelligence, but not as highly with tests of language development. This type of evidence of validity is called *concurrent validity* because different tests are given at the same time and the relationship between their scores is compared. If a new verbal test correlated highly with another test of verbal intelligence, this would be called evidence of *convergent validity* because the new test scores converge with scores from a known test of the same construct. If the new verbal test did not correlate as highly with a test of language proficiency, this would be evidence of *divergent validity* because the new test scores diverge with scores from a test which it is not supposed to relate to as highly. This shows that the two tests measure somewhat different constructs.

Many professionals ask us, "What is the validity coefficient for this test?" This is the wrong question to ask because validity is not a single number. It is a collection of evidence that supports the hypothesis that the test measures what it is supposed to measure. Some professionals ask, "Is this test valid?" Tests are not valid in general, but they are valid for specific purposes. A test of language development may be valid in assessing language development, for example, but not for assessing specific language disorders (e.g., pragmatic language disorder). So, the question should be, "Is this test valid for the purpose for which I intend to use it?" It is important to be clear about how you intend to use a test and then look for evidence of validity to support that use.



Clinical validity refers to how the test performs in specific clinical populations. A test of working memory, for example, might be expected to show much lower mean scores in a clinical sample of subjects known to have working memory disorder as compared to a nonclinical sample. In these studies, it is important that the clinical and nonclinical samples are matched according to other characteristics that may influence scores on the test, such as maternal education and age. In this way, you can be more certain that any differences observed between the clinical and nonclinical groups are truly due to the clinical disorder and not to other factors that were uncontrolled in the study because they are different between the two groups.

Another concept related to clinical validity is *statistical significance.* It is important for the score difference between the clinical and nonclinical groups to be statistically significant. It is even more important that the size of the difference is large enough to be clinically meaningful. Sometimes a difference of only a couple of points can be statistically significant, but the difference may not be large enough to be clinically useful.

Clinical validity refers to how the test performs in specific clinical populations.

To determine how meaningful the difference is, effect sizes are often reported in the test manual in a table, comparing a particular clinical group to a typically developing matched sample. Effect sizes of .20 are considered small, but perhaps still meaningful, depending on the purpose. Effect sizes of .50 and .80 are considered medium and large, respectively.

Sometimes a test has a cut score (or cut-off score) to determine if the client is at risk and should be referred for more in-depth testing or has a particular disorder. So, in a test of academic achievement, for example, one might say that any client with a score more than two standard deviations below the mean (i.e., 70) is classified as at risk for a learning disability, and any subject who scores above 70 is classified as not at risk for a learning disability. We want to see how well this cut score differentiates between the clinical and nonclinical samples. As shown in Table 1, subjects in the known clinical sample with scores below 70 are considered true positives because they are correctly classified as having the disorder. Subjects in the nonclinical sample with scores of 70 or higher are considered true negatives as they are correctly classified as not having the disorder.

Table 1. Sensitivity and Specificity

	< 70	≥ 70
Clinical	True positive	False negative
Non-clinical	False positive	True negative

Subjects in the known clinical sample with scores of 70 or higher are called *false negatives* because they have been classified as not having the disorder when they do have it. Those in the nonclinical sample with scores below 70 are called *false positives* because they have been incorrectly classified as having the disorder. False positives and negatives are always in a delicate balance, depending on where the cut score is set and the correct cut score depends on the purpose of testing. If the cut score is lowered, the percentage of false negatives increases and the percentage of false positives decreases. This may be appropriate in situations in which you want to be sure that you do not incorrectly label someone as having the disorder. If the cut score is raised, the percentage of false positives increase and the percentage of false negatives decrease. This is may be appropriate in situations when it is important to identify everyone who might have the disorder and incorrectly identifying a person does not have harmful consequences.

Some tests with well-developed cut scores do not require norms. This may be the case when the purpose of the test is to classify subjects as belonging to one or another group, but not to rate the severity of a disorder.

Test developers and researchers sometimes conduct studies with subjects already identified as having or not having a disorder. These studies are designed to evaluate the performance of a test. In real practice, you do not know ahead of time if the person you are testing has the disorder—after all, that is why you are testing. To determine how likely you are to correctly classify someone as having the disorder in real practice, divide the number of true positive cases in the sensitivity/specificity table by the sum of the number of true positive and false positives cases. This will give you an estimate of the *positive predictive power* of the test in applied situations. Even this method may have problems, however, if the percentage of clients in your practice with that disorder is much higher than in the study.

Conclusion

Standardized clinical assessments are extremely useful scientific instruments that inform, but do not replace, professional judgement. Clinical assessments are created for use and interpretation by trained professionals who also take into account the client's history, the referral question or concern, and other assessment data. We hope that this brief paper gives you an appreciation for the science behind clinical assessments and the basic knowledge to evaluate the quality of a testing instrument.

NCS Pearson. (2023). Standardized clinical assessment for practitioners: A primer.

866-335-8418 | PearsonClinical.ca

Pearson