

A Primer on Psychometrics

**The Important Points for
Speech Language-Pathologists**

Lawrence G. Weiss
Patricia Zureich

A Primer on Psychometrics

The Important Points for Speech Language-Pathologists

Lawrence G. Weiss, PhD, and Patricia Zureich, MA, CCC–SLP

Introduction

Psychometrics is a branch of statistics applied to the measurement of human behavior. Psychometricians use a specialized set of statistical tools to create scientifically valid assessments of various behaviors. Speech-language pathologists are particularly interested in assessments of language development and disorders, aphasia, autism, and early literacy. This overview of basic psychometric principles will serve as a means of helping you evaluate the quality of the assessment tools you use in your practice.

What distinguishes a standardized assessment from a nonstandardized assessment?

A standardized test might also be described as a “structured” test because it has a defined structure for administration procedures and scoring rules that are followed in the same manner by every professional who administers the test. Standardized assessments also have structured procedures for interpretation which usually involves comparing the client’s score to the scores of a representative sample of people with similar characteristics (e.g., age, sex, etc.). Typically, a test is considered to have been “standardized” if data have been collected on large numbers of subjects using a set of structured rules for administration and scoring. These data are used to determine the average score (mean) and the standard deviation, which the clinician then uses to benchmark the performance of the client being tested.

Nonstandardized tests are those that do not have data from a standardization sample in which the test was administered to a large number of subjects in exactly the same way by each clinician and scored according to a structured set of rules. Typically, these tests are created by individual clinicians who want to assess their clients and the tests are often shared with other clinicians who may adapt them for use with their own clients. Clinician-created measures are a step toward creating a standardized test because they assess constructs that are clinically meaningful. However, multiple problems can occur with clinician created measures when there is no scientifically supported basis for what constitutes a good or poor score, and when other clinicians administer the same test, but score it based on their own criteria. Sometimes individual clinicians must create a nonstandardized assessment because there are no standardized tests of the clinical construct of interest. Home-grown measures may or may not be a valid way to measure a skill—it is difficult to say until data has been collected to show that the test items and formats are effective in measuring a target skill. Not until data has been collected to verify the appropriateness of the administration procedures, the test items, and the scoring, can you have a reasonable degree of confidence that the test is suitable for the task.

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 were given 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. Psychometricians call these phenomena *item-context effects*. Psychometricians have also shown that in a testing situation, the relationship between the examiner and examinee can influence the level of effort the examinee puts into the test (an effect called motivation). In order to administer standardized assessments, examiners are trained to ask each client identical questions in a specific order and with 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 speech and language tests elicit responses from examinees that have to be judged for accuracy. Standardized scoring rules give examiners a common set of rules so that they can judge and score responses the same way.

Why are standardized tests an important part of clinical assessment practices?

Standardized assessments are an important part of clinical assessment practices in most of the helping professions because they help you gather and interpret data in standard manner, serve to confirm your clinical judgment, and support requests for services or reimbursement. In addition, many standardized tests are important to treatment because they can be used to guide interventions and measure treatment outcomes. Patterns of strengths and weaknesses documented by standardized assessments are often used to guide the development of an appropriate treatment plan. Repeat testing after a course of treatment can document patient progress. Finally, standardized tests can be useful in gathering a body of evidence about effective treatment 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 his or her 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 of a positive outcome. 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. Assuming the length of treatment was adequate, this would lead the outcomes-based therapist to consider another treatment.

Why do standardized tests have several different kinds of scores?

The number of items answered correctly in each subtest is called the subtest *raw score*. The raw score provides very little information to the clinician—you can only tell that the examinee got a few of the items correct or many of the items correct. This provides no information about how the score compares with the scores of other examinees of the same age and is usually converted into a standard score using a table developed using data collected during

standardization. A standard score is interpretable. “Standard scores” are standard because each raw score has been transformed to a predetermined value according to its position in the normal curve so that the mean (score) and the standard deviation (*SD*) are predetermined values (e.g., mean of 100 and *SD* of 15). Transforming the raw scores to predetermined values allows interpretation of the scores according to what we know about a normal distribution (normal curve).

There are different types of standard scores used in standardized tests. One of the most common is a metric in which the raw score mean of the test is transformed to a standard score of 100. Another very common type of standard score—the *T Score*—applies a score of 50 points to the raw score mean. Percentile ranks are also a type of score commonly used to interpret test results, and link directly to the standard scores based on the normal curve. A percentile rank indicates what percentage of people obtained that score or less. Thus, a percentile rank of 30 indicates that 30% of individuals in the standardization sample obtained that score or a lower score. Similarly, a percentile rank of 30 indicates that 70% of individuals in the standardization sample scored higher than that score.

Why must I convert the number of correct responses (raw score) into another score?

Raw scores need to be transformed into standard scores so that you can compare your client’s performance to the performances of other examinees of the same age or grade level. For example, let’s 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 gets 32 points (i.e., 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). Let’s say the average (or mean) score for second graders is 40 points. Now 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, and basically it tells us how much variability there is in the scores across 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 20 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.

By developing norms, psychometricians develop tables that you can use to convert each raw score (i.e., number of items correct) into a standard score for every age or grade covered by the test. When you look up your client’s raw score in the norms table to find the standard score, all of these statistical adjustments are already taken into account.

How do standard scores relate to the normal curve?

Standard scores are “standard” because the normative data (the original distribution of raw scores on which they are based) has 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 standard scores. As shown in the figure, 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 all of the individuals at his or her age or grade.

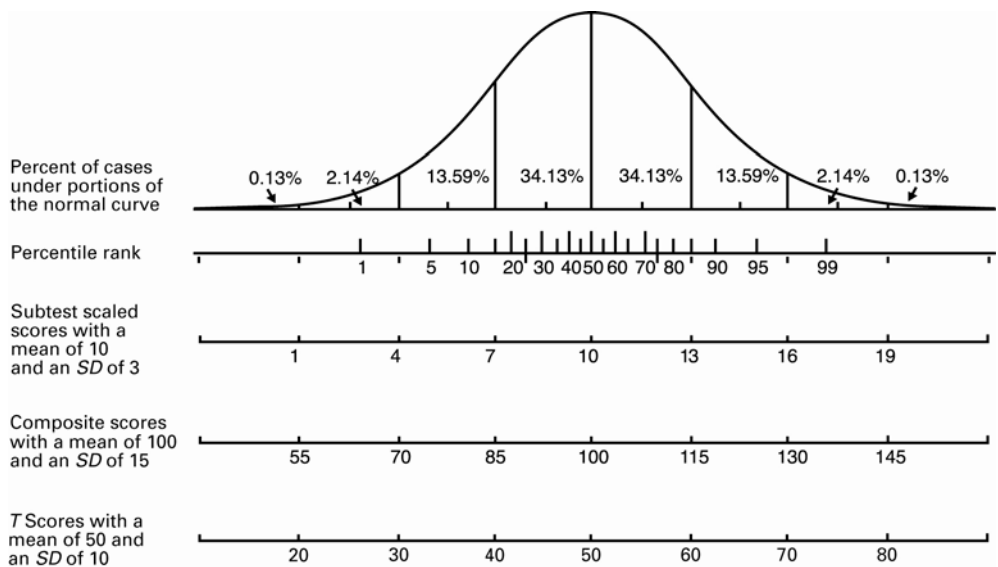
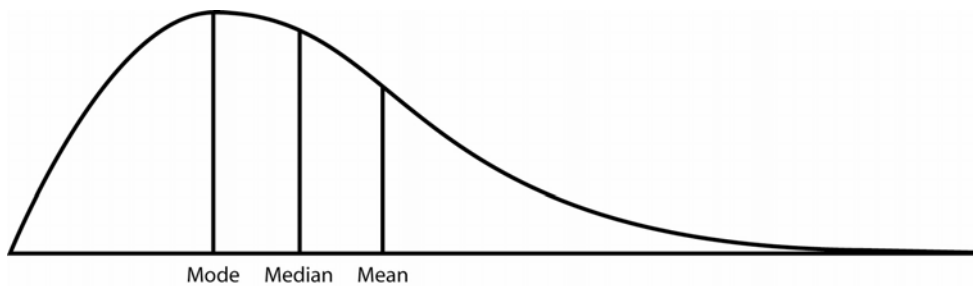


Figure 1. Normal Curve

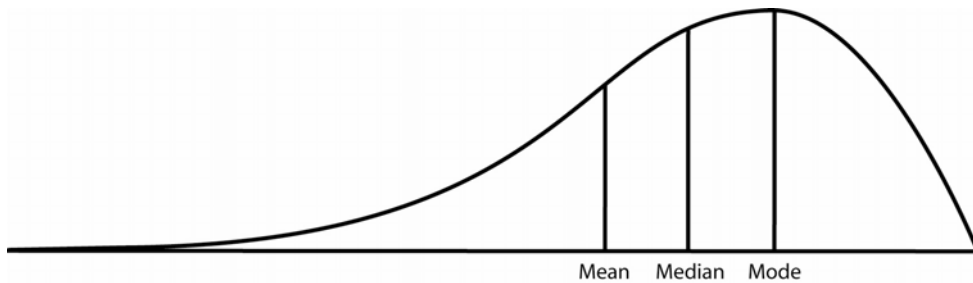
Standard scores are assigned to all raw scores based on the standard deviation. But, there are many different types of standard scores. Perhaps the most popular type of standard score is a metric where the mean is set to 100 and the standard deviation is set to 15. Recall in the previous example, the 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, so it would be assigned a standard score of 70 (i.e., 100–30). A standard score of 70 means the same thing for all tests normed using this 100/15 metric. Thus, the psychometricians have made it easy for you to interpret scores on different tests, and to compare scores across tests, without having to do these calculations on raw scores by yourself.

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).

For constructs that are normally distributed, the percentiles and standard deviations line up in the way described above (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 such 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. A negatively skewed distribution might be obtained when measuring a construct that most subjects of that age can easily perform and only very few cannot. For example, a test of phonological awareness for eight-year-olds might be negatively skewed because most eight-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 people cannot perform and only a few can. For example, a test of phonological awareness in three-year-old children may be positively skewed because most cannot perform these tasks, but a few can. In skewed distributions the percentiles and standard deviations units will not line up the same way as described on the normal curve. They will vary to the extent that the distribution is skewed.



Positively Skewed Distribution



Negatively Skewed Distribution

Figure 2. Positively and Negatively Skewed Distributions

Why do many tests have basal and ceiling rules?

Most tests used by SLPs are designed to assess 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 items on a test are arranged in order from easiest to hardest. *Basal rules* help you establish where to start the test so that you do not need to administer every item on the test. For example, if you are testing a 6-year-old girl for language development, the test developers might have you start the test with items appropriate for 5 ½-year-olds just to be sure that she understands the task and to give her some practice with the items. But, you would not need to administer items intended for 3- or 4-year-olds unless she had trouble responding to items designed for 5-year-old children. Typically, the *start point* in any test, subtest, or series of items is set at a level where 95% of all children that age have responded to the earlier items correctly. This helps reduce testing time and ensures only the items that are appropriate for each client are administered.

Ceiling rules help you know when to stop testing because you have exceeded the child's ability to respond correctly. Psychometricians analyze the standardization data to determine when you can be sure that if you administer another item the child will very likely get it wrong. Usually, the *stop rule* is set such that after a certain number of items are answered incorrectly there is less than a 5% chance that the examinee will be able to respond to any of the remaining items correctly. This reduces testing time, but equally importantly, prevents frustrating the examinee who may otherwise be administered many items that he or she cannot respond to correctly.

What are confidence intervals and why should I use them?

The precision and standard procedures used in administering and scoring standardized, norm-referenced tests may make you think that you can be 100% confident in the exact score obtained every time you administer the test. Unfortunately, because we are measuring human behavior and not a physical characteristic such as 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 examiners as discussed above.

For all these reasons, the client's true score may be slightly higher or lower than the specific score obtained by that examiner on that day. Thus, it is best to think of a range of scores that most likely describe your client rather than a single point score. This is why confidence intervals are used. The *confidence interval* is a range of scores around the client's obtained score that is sure to include the client's true score with 90% or 95% likelihood. Many tests report critical values that may be used to build confidence interval around each client's standard score, as plus and minus, (e.g., ± 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 a way of estimating the amount of 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 low. Psychometricians use the test's *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?

As you can see from the previous examples, the accuracy of any test’s standard scores depends on the accuracy of the raw score mean and standard deviation obtained from the normative sample that is used to create the transformations to standard scores. Thus, 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 by psychometricians 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.

But, 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 for which you use the test. However, this concept is sometimes misunderstood. A test of language development, for instance, does not have to include all or mostly all language disordered subjects in the normative sample. The performance of the language disordered subjects would pull the mean lower and the sample would no longer be representative of normal language development. Rather, the normative sample should include subjects with a language disorder in approximately the same percentage as they exist in the general population.

Other factors that are known from previous research to effect performance on the task of interest should also be represented in the normative sample. For example, it is known that mothers with less education tend to provide less than average language stimulation to their developing children and the lack of early language stimulation has a substantial impact on the child’s language development. Thus, when creating a test of language development it would be important to ensure that children from different parent education backgrounds are represented in approximately the same proportions as they are found in the general population. Similarly, regional variations in dialect may be important to represent in the normative samples of a test measuring syntax. 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.

Psychometricians must also be concerned with how long ago the normative sample was collected. Norms that were collected many years ago may no longer fairly represent today’s children or the current generation of adults. In state mandated achievement testing, there is a requirement to update norms every seven years. In the area of cognitive assessment, researchers have shown that norms tend to shift approximately 3 to 4 points every ten years. 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. Less is known about changes in language development across generations, but the older the norms, the more concern psychometricians have about their validity today.

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, tests of syntax, basic concepts, math, and phonological awareness.

What is reliability?

In general, *reliability* refers to the dependability of a test. Actually, there are several different types of reliability, and each type estimates a different source of possible measurement error. They all range between 0 and .99. The most common type is called *internal consistency reliability*, which assesses the extent to which all of 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 are correlated 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.

Another popular class of reliability is 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 and is sometimes referred to as the stability coefficient.

What is validity?

While internal consistency reliability is a way to determine if all of the items in a test measure the *same* thing, information is collected to provide evidence that the items measure the *right* thing. In other words, does a test of language development actually measure language development or is it really measuring verbal intelligence? To answer this question, one might design a study to show that a new language development test correlates highly with other established tests of language development but not as highly with tests of verbal intelligence. 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 compared. If the new language development test correlated highly with another test of language development, 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 language development test did not correlate as highly with a test of verbal intelligence, 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?" But, 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 valid for specific purposes. A test of language development may be valid in assessing language

development, for example, but not valid in assessing specific language disorders (e.g., pragmatic language disorder). So, one should ask, “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 the 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 receptive morphology, for example, might be expected to show much lower mean scores in a clinical sample of subjects known to have morphological disorder as compared to a nonclinical sample. (The term “nonclinical” simply means normal subjects.) 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/or age. In this way, you can be more certain that any differences observed between the groups are truly due to the clinical disorder and not to other factors that were uncontrolled in the study.

Another concept related to clinical validity is statistical significance. If a finding is statistically significant, it means that you are likely to be able to repeat the finding if you conduct the study again. It is important that the score difference between the clinical and nonclinical groups be statistically significant. But, it is even more important that the size of the difference be 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 clinically useful. To determine how meaningful the difference is, divide the difference by the standard deviation. Now you have a rough estimate of the *effect size*. Effect sizes (also called the Standard Difference) are often reported in the Examiner’s or Technical Manual in a table comparing a particular clinical group and 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-off score (or cut 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 our hypothetical test of language development one might say that any client with a score more than two standard deviations below the mean (i.e., 70) will be classified as having a language disorder, and any subject who scores above 70 will be classified as nonclinical. We want to see how well this cut score differentiates between the clinical and nonclinical samples. As shown in Figure 3, subjects in the known clinical sample with scores below 70 are considered true positives because they are correctly classified as having the disorder. This percentage represents an estimate of the *sensitivity* of the test (i.e., how sensitive the test is in identifying subjects with this disorder using this cut score). 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. This percentage represents an estimate of the *specificity* of the test (i.e., how specific the test is in ruling out subjects with this disorder using this cut score). These two figures combine to create the percent of correct classification, or *hit rate*.

	< 70	≥70
Clinical	True positive	False negative
Nonclinical	False positive	True negative

Figure 3. Sensitivity and Specificity

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 determining the correct cut score depends on the purpose of testing. If the cut score is moved lower, the percent of false negatives will increase and the percent of false positives will decrease. 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 moved higher, the percent of false positives will increase and the percent of false negatives will decrease. This may be appropriate in situations in which 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 has problems, however, if the base rate (prevalence or true frequency of occurrence) of people with the disorder in your work setting is much higher than in the study.

Conclusion

Standardized assessments are extremely useful tools that can help confirm your clinical judgment and guide your treatment plans. We hope that this brief paper gives you an appreciation for the science behind the tests you use in practice and, more importantly, the basic knowledge to evaluate the quality of the tests you use.