Chapter 8- Basic Statistics and Scoring Terminology Used in Assessment

As an educator, you will need to understand the scores that the various professionals of the multidisciplinary team report when they do their evaluations of children for a suspected disability. You may even be required to administer certain educational tests for a student. Therefore, it is essential that no matter what your role in the assessment process, you understand basic statistics and scoring terminology found in test manuals and used in assessment.

This section will provide you with the most frequently used terms used in assessment regarding test administration, statistics and scoring terminology.

The terms and corresponding examples will be presented in alphabetical order

A-Age Equivalent

An age equivalent is a very general score that is used to compare the performance of children at the same age with one another. It is the estimated age level that corresponds to a given score. Age equivalent scores are almost always given in years and months. For example, a child who gets an age equivalent score of 11-5 is performing as well as the average 11 year, 5 month old child.

B-Alternate Forms Reliability

Most standardized tests provide equivalent forms that can be used interchangeably. These alternate forms are typically matched in terms of content and difficulty. The correlation of scores on pairs of alternate forms for the same examinees provides another measure of consistency or reliability. Even with the best test and item specifications, each test would contain slightly different content and, as with test-retest reliability, maturation and learning may confound the results. However, the use of different items in the two forms conforms to our goal of including the extent to which item sets contribute to random errors in estimating test reliability.

C-Concurrent Validity

Concurrent validity is the extent to which a procedure correlates with the current behavior of subjects. It refers to how precisely a person’s present performance (e.g., a test score) estimates that person’s performance on the criterion measure administered at approximately the same time. In order to do a concurrent validity study, both measures must be given in close proximity. Normally, the administration of each of the two measures should not exceed more than two weeks. The procedure consists of administering the first instrument (i.e., the instrument to be validated) and very shortly thereafter, administering the criterion measure. Correlating the data from the two instruments then determines the concurrent validity (Overton).
D-Construct Validity

Construct validity seeks agreement between a theoretical concept and a specific measuring device or procedure. A classic question of construct validity involves intelligence tests, which determine intelligence by measuring subjects in areas such as vocabulary or problem-solving ability. The question of whether intelligence is being measured by these particular variables is an assessment of the test’s construct validity.

Construct validity can be broken down into two sub-categories: Convergent validity and discriminate validity. Convergent validity is the actual general agreement among ratings, gathered independently of one another, where measures should be theoretically related. Discriminate validity is the lack of a relationship among measures which theoretically should not be related.

E-Content Validity

Content validity refers to whether the individual items of a test represent what you actually want to assess. When we evaluate content validity, we are asking, “Does the content of our measure fairly and accurately reflect the content desired to be measured?” Thus, when we are measuring academic achievement with a new achievement test, we ask, “Is the score that we obtain truthfully measuring the actual academic achievement of the student?” Overall, content validity describes how well a test’s items reflect the area of learning to be assessed (Venn).

F-Correlation

A correlation is the amount of positive or negative relationship existing between two measures. For example, if the height and weight of a set of individuals were measured, it could be said that there is a positive correlation between height and weight if the data showed that larger weights tended to be paired with larger heights and smaller weights tended to be paired with smaller heights. The stronger those tendencies, the larger the measure of correlation.

G-Criterion Related Validity

Criterion related validity, also referred to as instrumental validity, is used to demonstrate the accuracy of a measure or procedure by comparing it with another measure or procedure which has been demonstrated to be valid. For example, imagine a hands-on driving test has been shown to be an accurate test of driving skills. By comparing the scores on the written driving test with the scores from the hands-on driving test, the written test can be validated by using a criterion related strategy in which the hands-on driving test is compared to the written test.
**H-Grade Equivalent**

This score represents the grade and month in school of students in the norm group whose test performance is equivalent to the test performance of a given student. For example, if a third-grade student obtains a grade equivalent of 4.8 on a mathematics test, it does not mean that the student has mastered all the mathematics that is taught in the school district during the first eight months of Grade 4. It means only that the student’s performance on this test is theoretically equivalent to the typical performance of students in the norm group who have completed eight months of Grade 4.

**I-Interrater Reliability**

Interrater reliability is the extent to which two or more individuals (coders or raters) agree. Interrater reliability addresses the consistency of the implementation of a rating system. Interrater reliability involves having two raters independently observe and record specified behaviors, such as hitting, crying, yelling, and getting out of the seat, during the same time period. For example, suppose two observers are to determine each time they see a certain child tap his pencil during a math lecture. Tapping the pencil during the math lecture is considered the target behavior. A target behavior is a specific behavior the observer is looking to record. After each observer determines the total number of times the target behavior occurs, the scores are compared, and an estimate of the percentage of agreement between the two observations is done (Venn, 2000). The reliability coefficient obtained in this case correlates the observations of two independent observers.

**J-Mean**

The mean is the arithmetic average of a set of numerical data. Statistically, the mean is represented by the symbol M. So, when we say that the mean on a test was 80%, we are stating that the “average” was 80%. The mean is simply calculated by adding up all the scores and then dividing the scores by the number of people taking the test.

Note: The mean is greatly affected by extreme scores. For example, suppose four students take an exam and receive scores of 90%, 95%, 100%, and 7%. The mean of the distribution is 73%. Notice though that three students did extremely well, but the one student who got a 7% took the mean from an A average to a C average.

**K-Median**

The median is defined as the score below which 50% of the cases fall. The median and the 50th percentile are the same. It is the middle score in a distribution. It is the point at which half the scores fall above and half the scores fall below. For example, the median value of the set \{5, 8, 9, 10, 11, 11, 13\} is 10.
L-Mode

The mode is the most frequently occurring score in a distribution. For example, the mode of the set \{13, 5, 9, 11, 11, 8, 10\} is 11. When you have two modes in a distribution, it is referred to as a bimodal distribution. If you have three or more modes in your distribution, it is referred to as a multimodal distribution.

M-Range

The range is the numerical difference between the largest and smallest values in a set of data. For example, the range of the set \{13, 5, 9, 11, 11, 8, 10\} is 8 (13-5 = 8).

Note: The range tells you nothing about the scores in between the high and low scores. And, if there is one extreme score, it can greatly affect the range. Suppose the distribution was 7, 9, 5, 9, 8, and 500. The range would be 495 (500 – 5 = 495). Yet, only one score is even close to 495 the 500.

N-Percentile Rank

A percentile is a value on a scale that indicates the percent of a distribution that is equal to it or below it. For example, a score at the 95th percentile is equal to or better than 95 percent of the scores. It is a score indicating the percentage of people or scores that occur at or below a given score. For example, if you have a percentile rank of 75 in a class, this means that you did as well as or better than 75% of the students in the class. A percentile rank of 16 means that you scored as well as or better than only 16% of the population. Percentile ranks range from the lowest (1st percentile) to the highest (99th percentile). A percentile rank of 83 means that a student has scored as well as or better than 83 percent of his peers on a test. Notice, however, it does not mean that she got a test score of 83%. The percentage correct on a test is not the same as the percentage of people scoring below a given score, the percentile rank. The 50th percentile normally signifies the average ranking or average performance.

There are two other types of percentiles used in assessment: quartiles and deciles. Quartiles divide scores into four units: 1–25, 26–50, 51–75, and 76–99. The first quartile (1–25) marks the lower quarter (Bottom 25%) or bottom fourth of all scores, whereas the fourth quartile represents the upper quarter (Top 25%). Deciles divide scores into tenths or 10 equal units. For example, the sixth decile is the point at which 60% of the scores fall below, whereas the ninth decile is the point at which 90% of the scores fall below.

In assessment, percentile ranks are very important because they indicate how well a child did when compared to the norms on a test. Knowing that a child had a percentile rank of 97 on a test would tell you that he is exceptional in this testing area, yet, knowing that he got a percentile rank of 7 would tell you that this is an area of weakness.
O-Predictive Validity

Predictive validity is the extent to which a procedure allows accurate predictions about a subject’s future behavior. It is a measure of a specific instrument’s ability to predict future performance on some other measure or criterion at a later date (Overton). For example, many colleges believe that the SAT has predictive validity with respect to how well a student will do in college. Similarly, the Graduate Record Exam is often required by admissions committees for graduate school because it is believed to have high predictive validity for future academic performance in graduate school.

P-Raw Scores

A student’s observed score on a test, i.e., the number correct. While raw scores do have some usefulness, they should not be used to make comparisons between performance on different tests, unless other information about the characteristics of the test is known.

When you administer any test, the first step in scoring almost always will be to calculate the number of correct items the student obtained. For example, if a student took a 20-question spelling test in your class, the first thing you would do is determine how many words the student spelled correctly. This score is known as the raw score. The raw score normally indicates the number of items correctly answered on a given test. In almost all cases, it is the first score a teacher obtains when interpreting data. A raw score is a test score that has not been weighted, transformed, or statistically manipulated.

In general, raw scores by themselves mean very little. For example, suppose the student in your class got 18 out of 20 correct on the spelling test. The number 18 has no real meaning. What is important is what you do with the 18. For example, most teachers would say the student got 18 out of 20 and turn it into a percentage indicating that the student got 90% (18/20 is 90%) on this test.

Q-Reliability

Reliability refers to the consistency of measurements. If a test lacks reliability, it is not stable, reproducible, predictable, dependable, meaningful, or accurate. In assessment, reliability relates to the confidence in an instrument to give the same score for a student if the test were given more than once. A reliable test produces similar scores across various conditions and situations, including different evaluators and testing environments (Venn).

R-Reliability Coefficients

The statistic for expressing reliability is the reliability coefficient. The reliability coefficient expresses the degree of consistency in the measurement of test scores. The symbol used to denote a reliability coefficient is the letter r with two identical subscripts (rxx). Reliability coefficients can range in value from 0.00 to 1.00. A reliability coefficient of rxx = 0.00 indicates absence of reliability, whereas a reliability coefficient of rxx = 1.00 demonstrates perfect reliability.
Acceptable reliability coefficients should never be below rxx = .90. A coefficient below rxx = .90 normally indicates inadequate reliability. A test should not be trusted if its reliability coefficient is low. High reliabilities are especially needed for tests used in individual assessment (Sattler). A reliability coefficient of rxx = .95 on a test means that 95% of a test score is accurate while only 5% consists of unexplained error. However, a test with a reliability coefficient of rxx = .60 does not have acceptable reliability because approximately 40% of the test score may be due to error (Venn).

S-Scaled Scores

Many tests used for assessment of children have subtests that comprise the entire test. For each subtest, a student receives a raw score. This raw score is often transformed into a scaled score. Scaled scores are very specific subtest scores. In many cases, scaled scores range from 1 to 19 with a mean of 10. They follow the following classification format:

<table>
<thead>
<tr>
<th>Scaled Score</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>Developmentally Delay</td>
</tr>
<tr>
<td>4–5</td>
<td>Well Below Average</td>
</tr>
<tr>
<td>6–7</td>
<td>Low Average</td>
</tr>
<tr>
<td>8–12</td>
<td>Average</td>
</tr>
<tr>
<td>13–14</td>
<td>High Average</td>
</tr>
<tr>
<td>15–16</td>
<td>Superior</td>
</tr>
<tr>
<td>17–19</td>
<td>Very Superior</td>
</tr>
</tbody>
</table>

For example, if a student gets only a scaled score of 7 on a Reading subtest but a 13 on a Math subtest, this indicates a much greater strength with respect to math than with reading as compared to the norms of his or her age group.

T-Split-Half Reliability or Internal Consistency

As the name suggests, split-half reliability is a coefficient obtained by dividing a test into halves, correlating the scores on each half, and then correcting for length (longer tests tend to be more reliable). The split can be based on odd versus even numbered items, randomly selecting items, or manually balancing content and difficulty. This approach has an advantage in that it only requires a single test administration. Its weakness is that the resultant coefficient will vary as a function of how the test was split. It is also not appropriate on tests in which speed is a factor (that is, where students' scores are influenced by how many items they reached in the allotted time).

U-Standard Deviation

The standard deviation is a statistic that indicates the amount of variability in a group of scores. When scores are normally distributed (i.e., when they are part of a bell-shaped, “normal” curve), about two-thirds of the scores are within one SD above and below the average (mean) score, and about 95% of scores are within 2 SDs of the mean. In almost any shaped distribution, all scores will be within 5 SDs of the mean score.
V-Standard Error of Measurement

Test manuals report a statistic called the standard error of measurement (SEM). It gives the margin of error that you should expect in an individual test score because of imperfect reliability of the test. The SEM represents the degree of confidence that a person’s “true” score lies within a particular range of scores. For example, an SEM of “2” indicates that a test taker’s “true” score probably lies within 2 points in either direction of the score he or she receives on the test. This means that if an individual receives a 91 on the test, there is a good chance that the person’s “true” score lies somewhere between 89 and 93 (America’s Learning Exchange, 2000, p.2). The SEM is a useful measure of the accuracy of individual test scores. The smaller the SEM, the more accurate the measurements. When evaluating the reliability coefficients of a test, it is important to review the explanations provided in the manual for the following:

W-Standard Scores

A standard score is a score that has been transformed to fit a normal curve, with a mean and standard deviation that remain the same across ages. Often, when doing assessment, you will have to tell parents and administrators the standard scores the child received on the given test and the appropriate classification that they represent. For some tests with a mean of 100 and a standard deviation of 15, the general classification system may appear as follows:

<table>
<thead>
<tr>
<th>Standard Score</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 70</td>
<td>Developmentally Delayed</td>
</tr>
<tr>
<td>70–79</td>
<td>Well Below Average or Borderline</td>
</tr>
<tr>
<td>80–89</td>
<td>Low Average</td>
</tr>
<tr>
<td>90–109</td>
<td>Average</td>
</tr>
<tr>
<td>110–119</td>
<td>High Average</td>
</tr>
<tr>
<td>120–129</td>
<td>Superior</td>
</tr>
<tr>
<td>130 and higher</td>
<td>Very Superior</td>
</tr>
</tbody>
</table>

Important Point: The above classification system is only one form of representing standard scores. Different tests may use different ranges and terminology.

X-Stanine

A stanine, an abbreviation for standard nines, is a type of standard score that has a mean of 5 and a standard deviation of 2. Stanine scores can range from 1 to 9. A stanine of 7 is 1 standard deviation above the mean (5 + 2). A stanine of 9 is 2 standard deviations above the mean (5 + 2 + 2). Conversely, a stanine of 3 is 1 standard deviation below the mean (5 – 2) and a stanine of 1 is two standard deviations below the mean (5 – 2 – 2).
Y-Test–Retest Reliability

A test-retest reliability coefficient is obtained by administering the same test twice and correlating the scores. In concept, it is an excellent measure of score consistency because it allows the direct measurement of consistency from administration to administration. This coefficient is not recommended in practice, however, because of its problems and limitations. It requires two administrations of the same test with the same group of individuals. This is expensive and not a good use of people's time. If the time interval is short, people may be overly consistent because they remember some of the questions and their responses. If the interval is long, then the results are confounded with learning and maturation, that is, changes in the persons themselves.

Z-T Score

A T score is another way to express test performance. T scores have a mean of 50 with a standard deviation of 10. Therefore, if you have a T score of 40 you are 1 standard deviation below the mean, whereas a T score of 60 would be 1 standard deviation above the mean.

AA-Validity

Validity refers to the issue of whether the test measures what it is intending to measure. Does a test of, say, mathematics ability measure that ability, or is reading comprehension a part of what is measured by the test? The validity of a test is constrained by its reliability. If a test does not consistently measure a construct or domain then it cannot expect to have high validity coefficients. The greater the validity of a test, the greater our confidence that it measures what it is designed to measure. Questions about validity are of ultimate importance for special educators because they address whether an instrument fulfills the function for which it was created. Accordingly, effort must be put into determining the validity of any measuring instrument that is to be used in a study. This section covers the most important and often utilized types of validity seen in special education assessment.

BB-z Scores

A z score indicates how many standard deviations a score is above or below the mean. A z score is a standard score distribution with a mean of zero and a standard deviation of one. For example, if a student has a z score of +1.0, this means that he scored 1 standard deviation above the mean on the test. If a student has a z score of –1.7, this means that he scored 1.7 standard deviations below the mean.