Week 4 -Content

Measurement Concepts in Test Administration and Interpretation

Important Characteristics of Assessments

Assessments in schools come in a variety of types and forms and are used for a variety of purposes. They may involve multiple choice items, constructed responses and observations of performance to name a few. Results of assessments may be used to plan instruction, to identify strengths and areas of need, to screen students, to monitor progress, to make diagnostic decisions, to make placement decisions, to evaluate programs, to predict success in future learning activities or settings, and to communicate performance to parents, educators and the community. Whatever the type of assessment employed or however the results are used, all assessments should possess the characteristics of validity, reliability and usability or practicality (Díaz-Rico & Weed, 2006; Linn & Gronlund, 2000).

Validity

"Validity is an evaluation of the adequacy and appropriateness of the interpretations and uses of assessment results." (Linn & Gronlund, 2000, p.73) Does a particular reading comprehension test tell us who has good reading comprehension skills, and who doesn't? Does the test measure what it intends to measure (BC Teachers' Federation, 2003)? It is important to make sure that assessments are used for the specific purpose and target population for which they were designed. If not, their use will not be valid. For example, if a test is designed to measure vocabulary at the elementary school level, then using its results as a measure for reading comprehension would not be valid. Administering that assessment to middle school students as a measure of vocabulary would not be valid either.

Reliability

Reliability refers to whether scores are consistent and dependable. An assessment is said to be reliable when its results are consistent and the level of measurement error is small. Assessment results cannot be expected to be totally consistent or reliable. Assessment results represent a measure of a sample of performance at a particular time.

Many factors other than what is being assessed may affect assessment results. For example, if a test is measuring writing skills, what factors other than students' writing skills may affect the results? Examples might include variations in effort, attention to task, familiarity with the test items and topics students are being asked to address, and who is scoring the test. To go a step further, would performance have differed if the student were assessed on a different day, with a different sample of items or if a different rater or teacher scored the test (Linn & Gronlund, 2000; BC Teachers' Federation, 2003)?

Reliability, or consistency of assessment results, is necessary for validity to be possible. An assessment that yields inconsistent (unreliable) results cannot produce valid information about what is being measured (e.g., phonics skills). However, high reliability of assessment results means that results are consistent, but does not necessarily mean you are measuring what you intended to measure or that you are using the results appropriately. For example, an assessment may produce consistent results but may not appropriately measure the subject area or domain it is intended to measure. (Linn & Gronlund, 2000). It is akin to measuring a person's blood pressure with an inaccurate sphygmomanometer or blood pressure monitor. You may get

consistent results over time, but they are not a valid or accurate measure of that person's blood pressure.

There are several kinds of reliability. Some of the most common types appear in the table below.

Types of Reliability	How to Measure
Test-Retest Reliability	• Obtained by comparing the results on the same assessment twice, separated by days, weeks or months. Reliability is the correlation between scores at time 1 and time 2.
Alternate Form Test Reliability	• Obtained by comparing the results obtained on equivalent or parallel forms of the same assessment administered at about the same time to the same individuals.
Internal Consistency	• Obtained by comparing one half of the test to the other half, or using methods such as Kuder-Richardson or Chronbach's Alpha reliability coefficients to identify the internal consistency of tests.

Adapted from: *Pinellas School District, & FCIT at USF (2007) Classroom Assessments available at http://fcat.usf.edu/assessment/basic/basicc.html*

Usability or Practicality

Usability of assessment procedures is an important practical consideration in selecting assessments. How long does the assessment take to administer? How easy or difficult are the administration procedures? What type of training and qualifications are needed to correctly administer the assessment? How easy or how difficult is it to interpret the assessment results and apply them to make informed decisions about students' strengths and needs? How expensive is the assessment? These are just a few of the questions that should be addressed when looking at the usability of assessment procedures (Díaz-Rico & Weed, 2006; Linn & Gronlund, 2000). Can you think of more?

Obtaining valid and reliable results should supercede usability considerations. Thus, it would not be appropriate to select an extremely short test which could substantially reduce the reliability of scores for the sake of expediency. Nor would it be appropriate to use a test that is easy to score that does not properly measure the content you are interested in measuring.

Comparing Norm-Referenced and Criterion-Referenced Tests

- Norm-referenced tests (NRTs) and criterion referenced tests (CRTs) are two major categories of tests used to measure and interpret student performance.

- NRTs and CRTs can both be standardized assessments. This means they are carefully constructed, field-tested and administered and scored in a standard and uniform way to all examinees across all settings. Standardization makes comparisons of student scores possible across students, classrooms and schools (Bond, 1996).

- NRTs and CRTs are similar in many other ways. For example, they both

    - require that the achievement domain to be measured be specified
    - use a sample of test items that is relevant and representative

- are judged by their validity, reliability and usability
- use the same kinds of test items and the same rules for item writing (Linn & Gronlund, 2000).

- However, the purposes of NRTs and CRTs are different. The main purpose of a NRT is generally to compare a student's performance to that of a norm group that is comprised of a large national sample of students at the same grade and/or age level. NRTs rank test takers or compare them to each other in terms of performance. On the other hand, the main purpose of a CRT is to identify how well test takers have learned what they are expected to know and be able to do according to a specified set of standards or outcomes (Bond, 1996; Linn & Gronlund, 2000).

- The FCAT-NRT and the Stanford Achievement Test are examples of NRTs. They allow for comparisons of the reading and mathematics achievement of our students to that of students across the United States.

- The FCAT-Sunshine State Standards (FCAT-SSS) is an example of a CRT. It assesses student's mastery of Florida's SSS benchmarks in reading, mathematics, science and writing (Florida Department of Education, 2007).

Understanding Measurement Terms Used in Interpretation of Test Results-

Common terms used in test interpretation appear next.

Raw Score - This is the score achieved on a test without any manipulations. The raw score on a test would be the number of items correct or the number of points earned. If a student got 20 items correct on a 25 item test, his raw score would be 20.

Percent Correct - This is the number of points a student has earned divided by the number of possible points. For the student who answered 20 items correctly on a 25 item science test where each item is weighted the same, the percent correct would be 80%. This is calculated as follows: (20/25) X 100 = 80% correct.

Mean - This is an average used to represent all scores in a distribution of scores. You calculate the mean by adding all the raw scores in a group of scores and dividing by the number of scores. For example, the scores on a science test in class A are as follows:

10 10 12 12 12 14 14 15 15 16 16 18 18 19 19

The mean is then

$$\text{Mean} = \frac{10+10+12+12+12+14+14+15+15+16+16+18+18+19+19}{15} = \frac{220}{15} = 14.7$$

Measurement Concepts in Test Administration and Interpretation

Median - This is another average score used to represent all the scores in a group of scores. It is found by: a) placing in rank order all the scores that pertain to the distribution, and b) identifying the middle score, that is, the score that divides the distribution of scores in half.

Fifty percent of the scores can be found above the median and fifty percent of them can be found below the median. The median is the 50th percentile in a distribution of scores.

For example, the median score identified for the same science test scores of students in class A referred to previously is 15. There are seven scores above it and seven scores below it.

10 10 12 12 12 14 14 **15** 15 16 16 18 18 19 19

Median

Standard deviation - This value indicates how different the scores in a distribution are from the mean score.

In a distribution where most scores are close in value around the mean, the value of the standard deviation will be smaller than in another distribution where the scores are scattered widely around the mean.

This can be illustrated as follows:

Given the distribution of scores on the science test for class A is,

10 10 12 12 12 14 14 15 15 16 16 18 18 19 19

the value of the mean is **14.67** and the standard deviation is **3.04**.

Given the distribution of scores on the same science test obtained by students in class B is,

13 13 14 14 14 14 15 15 15 15 15 15 16 16 16

the value of the mean is also **14.67**, but in this case the standard deviation is **0.98**.

The importance of the standard deviation is that it can identify whether a group's performance is heterogeneous (varied for all students in the group, as in the class A example), or homogeneous (similar for all students in the group, as illustrated in the class B example).

Standard Score - Also known as z-score value, the standard score indicates how far away from the mean a certain score is in standard deviation units. For example, when Sally's raw score of 18 is transformed into a z-score (or standard score), it would yield the following z-score value:

$$\text{Standard score (z-score)} = \frac{\text{raw score - mean}}{\text{standard deviation}} = \frac{18 - 14.7}{3.04} = 1.09 \text{ S.D.}$$

Interpretation - Sally's score of 18 on this science test is 1.09 standard deviations above the mean of her group.

Percentile Scores or Percentile Ranks - Not to be confused with percent correct, percentile ranks indicate where a student's score is relative to the scores of other members of his or her group. That is, for a particular student's percentile score, a percentile rank is the percentage of the group that achieved scores at or below that percentile score.

> For example, if Sally achieved a percentile rank score of 87 on the FCAT-NRT Reading, this indicates that 87 percent of the other students in the same grade who were part of the norming group scored at or below Sally or to put it another way, Sally's score was above 87 percent of the scores of other students in the same grade who were part of the norming group.

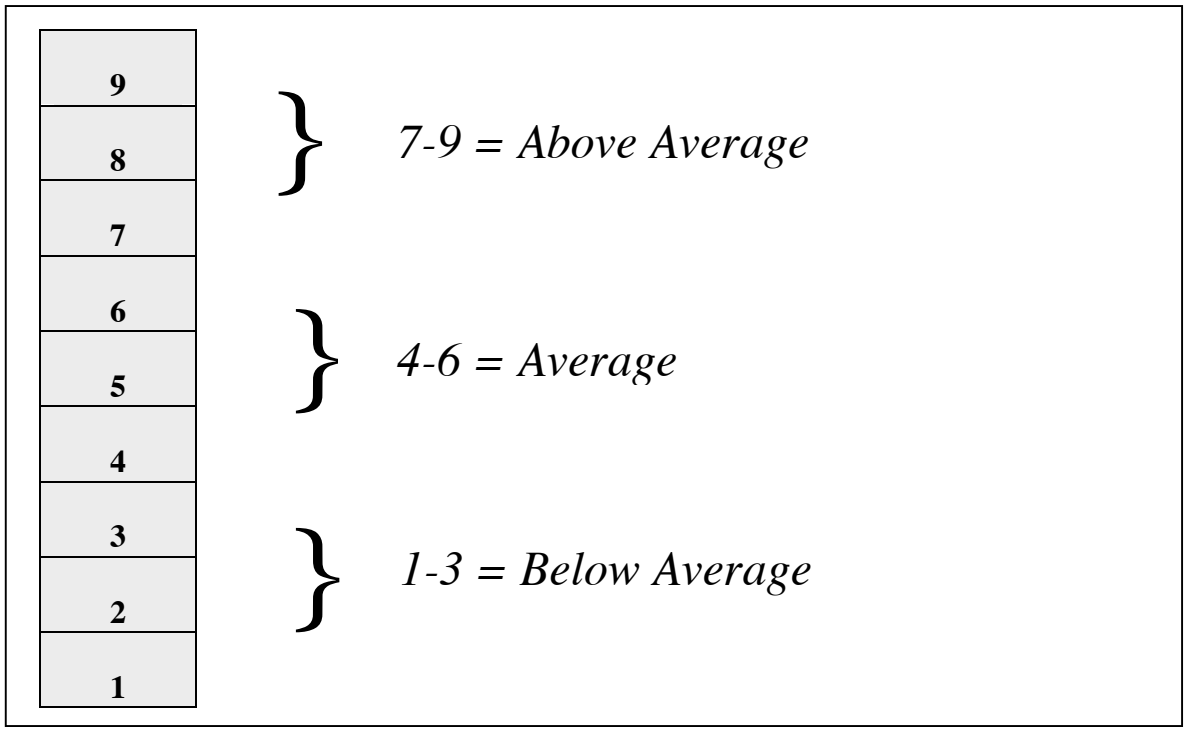
Percentile ranks range from 1 to 99. A percentile rank of 50 is average.

Normal Curve Equivalent (NCE) - NCEs are standard scores with a mean of 50 and a standard deviation of 21.06. NCEs range from 1 to 99. Unlike percentile ranks, NCE scores can be averaged and used for further statistical calculations. For this reason, they are used mostly when establishing comparisons of group performance.

Stanines - These types of scores also indicate the relative position of a score in reference to a group. Stanines are achieved by transforming the standard scores on a test into a scale with a mean of 5 and a standard deviation of 2. This transformation results in Stanine values of 1 to 9.

Measurement Concepts in Test Administration and Interpretation

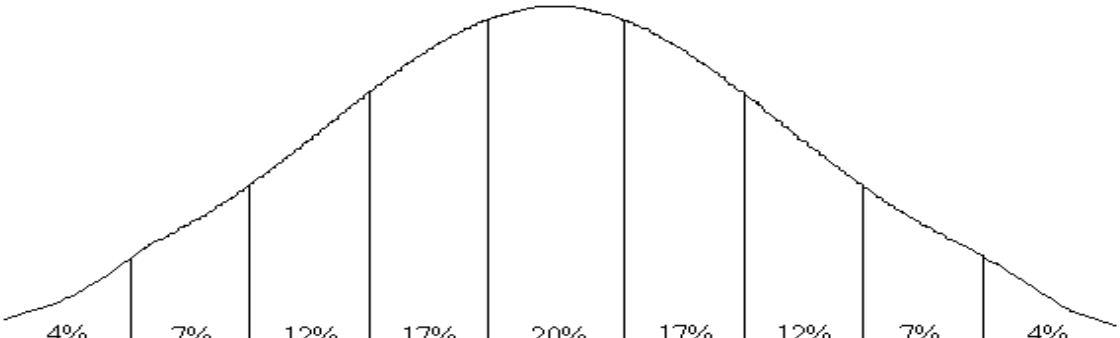
Illustration of Stanines



For example, Sally's percentile rank score of 87 on the FCAT-NRT Reading would be equal to a stanine of 7.

The figure below illustrates how percentile ranks, NCE scores and stanines are related in a normal curve distribution. For example, a stanine of 5 would be comparable to a percentile rank of 40-59 and a NCE of 44.7 - 54.8.

Relationship of Percentile Ranks, NCEs and Stanines in a Normal Curve Distribution

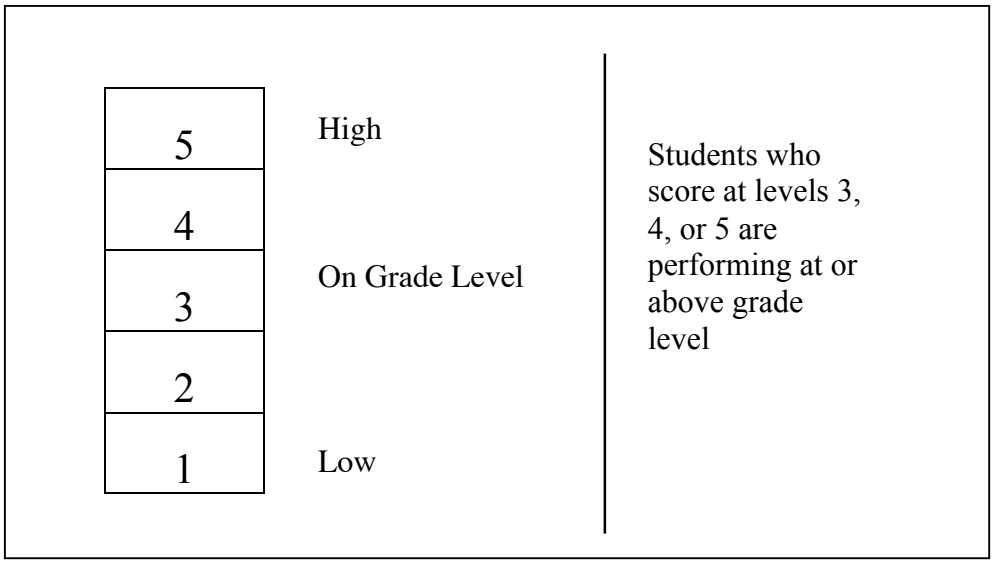


4%	7%	12%	17%	20%	17%	12%	7%	4%
				Stanines				
1	2	3	4	5	6	7	8	9
				Percentile Ranks (PR)				
Below 4	4 - 10	11 - 22	23 - 39	40 - 59	60 - 76	77 - 88	89 - 95	Above 95
				Normal Curve Equivalents (NCE)				
1.0 – 10.4	13.1 – 23.0	24.2 – 33.7	34.4 – 44.1	44.7 – 54.8	55.3 – 64.9	65.6 – 74.7	75.8 – 84.6	86.9 – 99.0

Measurement Concepts in Test Administration and Interpretation

Scale Scores - These are also transformed or converted scores used to report results on an entire test. For example, on the FCAT-SSS scale scores range from 100-500 for each content area and grade level. (Florida Department of Education, 2007).

Achievement Levels on FCAT-SSS - Achievement levels describe a test taker's success on the SSS tested on the FCAT. Achievement levels range from a low of 1 to a high of 5. (Florida Department of Education, 2007).



There are specific scale scores and developmental scale scores associated with each achievement level by grade and by content area.

Developmental Scale Scores - This is a type of scale score used on the FCAT-SSS to determine a student's yearly progress from one grade to the next. The Developmental Scale Score is called the FCAT score on the student and parent report.

Measurement Concepts in Test Administration and Interpretation

References

BC Teachers; Federsation (2003). *A primer on educational data.* Retrieved June 8, 2007, from http://bctf.ca/IssuesInEducation.aspx?id=5722&printPage=true.

Bond, L.A. (1996). Norm- and criterion- referenced testing. *Practical Assessment, Research & Evaluation,* 5 (2). Retrieved May 23, 2007 from http://PAREonline.net/getvn.asp?v=5&n=2.

Díaz-Rico, L. T. & Weed, K. Z. (2006). *The crosscultural, language, and academic development handbook: A complete K-12 reference guide* (3rd Ed.). Boston, MA: Pearson Education, Inc.

Florida Department of Education (2007). *Understanding FCAT reports 2007.* Tallahassee, FL: author. Retrieved June 4, 2007, from http://fcat.fldoe.org/fcatUnderstandReports.asp.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th Ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Pinellas School District and FCIT at University of South Florida. *Classroom assessment basic concepts: Reliability and validity*. Retrieved April 12, 2007 from http://fcit.usf.edu/assessment/basic/basicc.html.