TEST SCORES: A GUIDE TO UNDERSTANDING AND USING TEST RESULTS

By Dawn P. Flanagan, PhD, & Lenny F. Caltabiano
St. John's University

NATIONAL ASSOCIATION OF SCHOOL PSYCHOLOGISTS

When a student takes either an individually or group-administered standardized test at school, the results are made available to both parents and teachers. It is important that parents and teachers understand the meaning of scores that come from standardized tests. This handout provides a description of common terms used to describe test performance. You are also encouraged to refer to handouts on Psychological Reports (Flanagan & Caltabiano) and Intellectual Assessment (Ortiz & Lella) to gain a better understanding of the evaluation process (See "Resources").

Frequently Used Terms

The results of most psychological tests are reported using either *standard scores* or *percentiles.* Standard scores and percentiles describe how a student performed on a test compared to a representative sample of students of the same age from the general population. This comparison sample or group is called a *norm group.* Because educational and psychological tests do not measure abilities and traits perfectly, standard scores are usually reported with a corresponding *confidence interval* to account for error in measurement.

Standard Score

Most educational and psychological tests provide standard scores that are based on a scale that has a statistical mean (or average score) of 100. If a student earns a standard score that is less than 100, then that student is said to have performed below the mean, and if a student earns a standard score that is greater than 100, then that student is said to have performed above the mean. However, there is a wide range of average scores, from low average to high average, with most students earning standard scores on educational and psychological tests that fall in the range of 85–115. This is the range in which 68% of the general population performs and, therefore, is considered the *normal limits* of functioning.

Classifying standard scores. However, the normal limits of functioning encompass three classification categories: *low average* (standard scores of 80–89), *average* (standard scores of 90–109), and *high average* (110–119). These classifications are used typically by school psychologists and other assessment specialists to describe a student's ability compared to same-age peers from the general population.

Subtest scores. Many psychological tests are composed of multiple *subtests* that have a mean of 10, 50, or 100. Subtests are relatively short tests that measure specific abilities, such as vocabulary, general knowledge, or short-term auditory memory. Two or more subtest scores that reflect different aspects of the same broad ability (such as broad Verbal Ability) are usually combined into a *composite* or *index* score that has a mean of 100. For example, a Vocabulary subtest score, a Comprehension subtest score, and a General Information subtest score (the three subtest scores that reflect different aspects of Verbal Ability) may be combined to form a broad Verbal Comprehension Index score. Composite scores, such as IQ scores, Index scores, and Cluster scores, are more reliable and valid than individual subtest scores. Therefore, when a student's performance demonstrates relatively uniform ability across subtests that measure different aspects of the same broad ability (the Vocabulary, Comprehension, and General Information subtest scores are both average), then the most reliable and valid score is the composite score (Verbal Comprehension Index in this example). However, when a student's performance demonstrates uneven ability across subtests that measure different aspects of the same broad ability (the Vocabulary score is below average, the Comprehension score is below average, and the General Information score is high average), then the Verbal Comprehension Index may not provide an accurate estimate of verbal ability. In this situation, the student's verbal ability may be best understood by

looking at what each subtest measures. In sum, it is important to remember that unless performance is relatively uniform on the subtests that make up a particular broad ability domain (such as Verbal Ability), then the overall score (in this case the Verbal Comprehension Index) may be a misleading estimate.

Percentile

Standard scores may also be reported with a percentile to aid in understanding performance. A *percentile* indicates the percentage of individuals in the norm group that scored below a particular score. For example, a student who earned a standard score of 100 performed at the 50th percentile. This means that the student performed as well as or better than 50% of same-age peers from the general population. A standard score of 90 has a percentile rank of 25. A student who is reported to be at the 25th percentile performed as well or better than 25% of same-age peers, just as a student who is reported to be at the 75th percentile performed as well or better than 75% of students of the same age. While the standard score of 90 is below the statistical mean of 100 and is at the 25th percentile, this performance is still within the *average* range and generally does not indicate any need for concern.

Confidence Interval

Psychological tests do not measure ability perfectly. No matter how carefully a test is developed, it will always contain some form of *error* or unreliability. This error may exist for various reasons that are not always readily identifiable. In order to account for this error, standard scores are often reported with confidence intervals.

Confidence intervals represent a range of standard scores in which the student's true score is likely to fall a certain percentage of the time. Most confidence intervals are set at 95%, meaning that a student's true score is likely to fall between the upper and lower limits of the confidence interval 95 out of 100 times (or 95% of the time). For example, if a student earned a standard score of 90 with a confidence interval of +5, this means that the lower limit of the confidence interval is 85 (that is, 90 – 5 = 85) and the upper limit of the confidence interval is 95 (90 + 5 = 95). The standard score of 90 may be reported in a psychological report as 90 + 5 or 90 (85 – 95). Although the student's score on the day of the evaluation was 90 in this example, the *true score* may be lower or higher than 90 owing to an error associated with the method in which the ability was measured. Therefore, it is more accurate to say that there is a 95% chance that the student's true performance on this test falls somewhere between 85 and 95.

Tests that are highly reliable have relatively small confidence bands associated with their scores, indicating that these tests provide the most consistent scores across time.

Example: Reporting Scores

The following statement is one that can be commonly found in a psychological report and can be used to illustrate these definitions: "Jacob obtained a *standard score* of 93 + 7 on a test of reading comprehension, which is ranked at the *33rd percentile* and is classified as *average*." This is what that statement means: First, Jacob's observed score fell below the mean of 100. Second, Jacob did as well as or better than 33% of students his age from the general population. Third, there is a 95% chance that Jacob's true score falls somewhere between 86 and 100. Fourth, Jacob's performance is considered *average* relative to same-age peers from the general population. The table at the end of this handout provides commonly used performance classifications for standard scores and percentiles.

Understanding the Assessment Report

Type of norms used. It is important to take note of the types of norms used when reading test results in a psychological or school assessment report. A student's performance on a standardized test can be compared to other students of the same age (age norms) or of the same grade (grade norms). Age norms are always used for tests of intellectual ability so that comparisons can be made to same-age peers. The use of grade norms is related to the type of test being utilized or may be dictated by certain situations. For example, grade norms may be most appropriate for achievement tests when a student has repeated a grade and to see how the student's performance compares to grade-level peers.

Use of age or grade equivalents. Age and grade equivalents are different from age and grade norms. Essentially, the age and grade equivalents are scores that indicate the typical age or grade level of students who obtain a given score. For example, if Jacob's performance on the test of reading comprehension is equal to an age equivalent of 8.7 years and a grade equivalent of 2.6, this means that his obtained *raw score* is equivalent to the same number of items correct that is average for all 8-year, 7-month old children included in the norm group on that particular reading comprehension test. Additionally Jacob's score is equivalent to the average reading comprehension performance of all children included in the normative sample who were in the sixth month of second grade. The age or grade equivalents do *not* mean that Jacob is

functioning on an 8-year-old, mid-second grade level. Remember that Jacob's standard score of 93 is classified as *average* and falls within the normal range of functioning. Consequently, it is always important to make decisions and interpretations about *normal* functioning using standard scores, not age and grade equivalents.

Validity of scores. Reports of assessment results typically include a statement as to the validity—or accuracy—of the test scores. There are many factors that can influence a student's test performance. These factors may include, but are not limited to, behavior during testing, the presence of distractions during testing, the student's cultural and linguistic background, and the student's physical health at the time of testing. An educational or psychological test report should indicate whether any of these factors were present and how they may have affected the results of the test, thereby compromising the validity of the findings. Typically, this information, appearing in the Behavioral Observations section of a psychological report, aids in assessing the validity and usefulness of the test findings. If the school psychologist did not observe any unusual behaviors during testing and if no other factors, internal (lack of motivation, depressed mood, fatigue) or external (loud voices outside the testing room), were believed to have had an adverse affect on test performance, then the psychologist's statement about the validity of the findings may be like this: "Overall, the current test results appear to represent a valid estimate of Jacob's cognitive and academic functioning." This statement assists the reader in determining whether the results from the psychological tests administered to the student may be used confidently to make diagnostic and educational decisions.

Summary

When parents and teachers better understand the meaning of scores from educational or psychological evaluations, they are able to better plan to meet student needs. Additional information is available in the "Resources" below, and from the assessment professionals at your school, such as the school psychologist or counselor.

Resources

Flanagan, D., & Caltabiano, L. (2004). Psychological reports: A guide for parents and teachers. In A. Canter, L. Paige, M. Roth, I. Romero, & S. Carroll (Eds.), *Helping children and home and school II: Handouts for families and educators.* Bethesda, MD: National Association of School Psychologists.

Harcourt Assessment (n.d.). *Some things parents should know about testing.* Available: *http://marketplace.psychcorp.com* (See Resource Center, About Testing)

Ortiz, S., & Lella, S. (2004). Intellectual assessment and cognitive abilities: Basics for parents and educators. In A. Canter, L. Paige, M. Roth, I. Romero, & S. Carroll (Eds.), *Helping children and home and school II: Handouts for families and educators.* Bethesda, MD: National Association of School Psychologists.

Wright, P. D., & Wright, P. D (2000). *Understanding tests and measurement for the parent and advocate.* Available: *www.ldonline.org/ld_indepth/assessment/tests_measurements.html*

References for the Table

Flanagan, D., & Ortiz, S. (2001). *Essentials of cross-battery assessment.* New York: Wiley.

Flanagan, D., Ortiz, S., Alfonso, V., & Moscolo, J. (2002). *The achievement test desk reference: Comprehensive assessment and learning disabilities.* Boston: Pearson, Allyn & Bacon.

Woodcock, R. W., & Mather, N. (1989). WJ—R Test of Cognitive Ability—Standard and Supplemental Batteries: Examiner's manual. In R. W. Woodcock & M. B. Johnson (Eds.), *Woodcock-Johnson Psycho-Education Battery—Revised.* Chicago: Riverside.

Websites

Harcourt Assessment—
http://marketplace.psychcorp.com

Dawn P. Flanagan, PhD, is a Professor and Coordinator of the School Psychology program at St. John's University, Jamaica, NY. Lenny F. Caltabiano is a doctoral student in school psychology at St. John's University.

Classifying Test Scores

Result		Classification of Performance	
Standard score range	Percentile rank range	Descriptive	Normative
>131	98–99+	Very superior	Normative strength; 16% of the population
121–130	92–97	Superior	
116–120	85–97	Above average	
111–115	76–84	High average	Normal limits; 68% of the population
90–110	25–75	Average	
85–89	16–24	Low average	
80–84	9–15	Below average	Normative weakness; 16% of the population
70–79	3–8	Deficient	
< 69	< 2	Very deficient	

Note. Classifications are based on those described in Flanagan and Ortiz (2001) and Flanagan, Ortiz, Alfonso, and Mascolo (2002) and were adapted from Woodcock and Mather (1989)