Norm-Referenced Testing

DERIVED SCORES

REVIEWING THE ADEQUACY OF A NORM GROUP

STRENGTHS AND WEAKNESSES OF NORM-REFERENCED TESTS

THE STATUS OF NORM-REFERENCED TESTING

Norm-referenced testing is integral to the practice of psychological and educational testing. Originated in the work of modern statistics, this assessment method assumes that human traits and characteristics, such as intelligence, academic achievement, and behavior, are distributed along a normal probability or bell-shaped curve (hereafter, referred to as the normal curve). The normal curve represents the norm or average performance of a population and the scores that are above and below the average within that population. The norms for a test include percentile ranks, standard scores, and other statistics for the norm group on which the test was standardized. A certain percentage of the norm group falls within various ranges along the normal curve. Depending on the range within which test scores fall, scores correspond to various descriptors ranging from deficient to superior.

Figure 1 ILLUSTRATION BY GGS INFORMATION SERVICES. CENGAGE LEARNING, GALE.

A Norm-Referenced Test (NRT) compares an examinee's test performance to those of the examinee's same-age peers from the test's norm group. This comparison permits a more meaningful interpretation of the individual's score. An examinee's test score is compared to that of a norm group by converting the examinee's raw scores into derived or scale scores. As shown in Figure 1, derived scores correspond to the normal curve, thus providing an interpretive framework for examinees administered the NRT. The norm group can consist of a larger population, such as a representative population of children from the United States (i.e., a national norm group), or it can consist of a smaller, more limited population, such as all children in an individual school or school district (i.e., a local norm group).
The conversion of examinees' raw scores into derived scores provides a system of common metrics that facilitate test interpretation. There are multiple derived scores reported for NRTs; however, the more common ones used for interpretive purposes are standard scores, standard deviations, scale scores, T-scores, percentile ranks, age-equivalent scores, and grade-equivalent scores. A standard score has a statistical mean or average of 100 and conveys how far an examinee's test score varies or deviates from the average of the distribution. The extent to which a score varies or deviates from the average is expressed as a standard deviation. For example, a standard score of 115 on a NRT, such as an intelligence quotient (IQ) test, is one standard deviation above the mean of 100 and falls within the High Average range.

Scale scores yield information about an examinee's performance on a sub-domain or subtest of a NRT. These scores have a mean or average of 10 and a standard deviation of 3 points. Therefore, a scale score of 7 on a NRT subtest indicates that the examinee's skill, as measured by the subtest, is one standard deviation below the mean of 10 and corresponds to the Low Average range. T-scores are a different distribution of standard scores in that the mean of the distribution is 50 with a standard deviation of 10 points. NRTs assessing behavior typically report T-scores. Consequently, a T-score of 60 on a behavior rating scale is one standard deviation above the mean of 50 and corresponds to the High Average/At-Risk range. Percentile ranks indicate an examinee's position relative to the norm group. A percentile rank is a point in the distribution at or below which a certain percentage of scores falls. A child obtaining a percentile rank of 85 on an intelligence test indicates that the child's performance is equal to or greater than 85 percent of the child's same-age peers in the test's norm group. It does not mean the child obtained 85 percent of items correct on the NRT.

Grade- and age-equivalent scores are two of the most commonly used scores to report test results, yet they are also commonly misunderstood. These scores indicate that a student has attained the same raw scores (or number of items correct) as the average student of a certain grade or age level in the test's norm group. For example, if a student obtains a grade-equivalent score of 4.5 on a test of basic reading skills, this means the student obtained the same raw score as the average student in the fifth month of the fourth grade in the test's norm group. It does not mean the student acquired or demonstrated the same level of proficiency consistent with curricular expectations as the average student in the fifth month of the fourth grade at the student's school. Likewise, if a student obtains an age-equivalent score of 8.0 years on a test of basic reading skills, this means the student obtained the same raw score as the average 8-year-old student in the test's norm group. It does not mean the student acquired or demonstrated the same level of proficiency consistent with curricular expectations as the average 8-year-old student in the student's school. Consequently, grade and age-equivalent scores should not be interpreted literally. It is critical that test administrators and consumers understand the meaning and interpretation of these and other derived scores, as misinterpretation can lead to serious consequences for the examinee, including possible misdiagnosis, misclassification, and/or inappropriate educational placement and services. Sattler and Lyman provide further, extensive information about these and other derived scores used for testing interpretive purposes.

Of note is the fact that NRTs are imperfect by nature. The scores yielded by NRTs are referred to as observed scores versus absolute or true scores. An observed score is one attained by an examinee, whereas a true score is one that is error-free and hypothetical in nature. For this reason, norm-referenced standard test scores have bands of error, which are expressed as either 90 percent or 95 percent confidence intervals. The 90 percent confidence interval indicates the range of standard scores within which an examinee's true score would fall 90 out of 100 times on repeated
assessments. Likewise, the 95 percent confidence interval, a more conservative estimate, indicates the standard score range within which an examinee's true score would fall 95 out of 100 times on repeated assessments.

REVIEWING THE ADEQUACY OF A NORM GROUP

It is necessary to examine key characteristics of a test's norm group to ensure the adequacy of the norms, hence, the appropriateness of the test. Manuals that come with commercially developed tests should provide this information. According to Salvia and Ysseldyke, some of the important characteristics of a test's norm group are: a) the representativeness of the group, b) the number of individuals in the group, and c) the relationship of the norms to the purpose of testing. Adequate representation is dependent, in part, upon the demographic characteristics of the individuals in the norm sample, including their age, sex, race/ethnicity, parent education level, and geographic location. It is important that the sample of individuals in the norm group be proportioned across the aforementioned variables according to their prevalence in the reference population.

For example, the norm group of an intelligence test developed in the United States for children ages 6 years, 0 months to 16 years, 11 months, should include representative proportions of children within this age range according to selected demographic variables based on U.S. Bureau of the Census data. The age of the norms (i.e., the difference in time between the year in which the norm group was administered the test and the year an examinee is administered the test) is also a critical dimension when evaluating the representational aspect. In order for a norm group to be representative, it must be current. Reschly suggests that test norms older than 10 to 12 years may lead to inflated test scores, based on research indicating that intelligence in the general population increases at a rate of approximately three points per decade. The test consumer should be aware that the older the norm group is, hence, the older the test norms are, the less accurately the group and norms represent the current reference population.

The number of individuals comprising the norm group is also important because a large norm group assures reliability of the test as well as representation of outliers in the reference population. Salvia and Ysseldyke recommend that the norm group should contain at least 100 subjects per age or grade level. Also of crucial significance is the relationship of the norms to the purpose of testing. For example, if the purpose of testing is to determine how a child performed on a school district reading assessment, the norms developed for the district for that particular test administration would be the most appropriate reference of comparison. If the purpose is to ascertain how a child is performing intellectually, then the norms of a nationally standardized test of intelligence would be the most appropriate reference of comparison.

Finally, norms that include the performance of individuals with special needs are an important consideration for accurate representation, particularly when evaluating individuals for eligibility for special education services and program placement. To the extent that test norms are adequate, they allow for meaningful comparisons and accurate information about the population. The Standards for Educational and Psychological Testing (AERA et al., 1999) states: “Norms, if used, should refer to clearly described populations. These populations should include individuals or groups to whom test users will ordinarily wish to compare their own examinees” (p. 55). To this end, test users must determine the applicability of a test to any given individual or group.

STRENGTHS AND WEAKNESSES OF NORM-REFERENCED TESTS

Ornstein describes a number of strengths of NRTs, including but not limited to the following: a) they assume statistical rigor in that they are reliable (i.e., dependable and stable) and valid (i.e., measure
what they are reported to measure); b) the quality of test items is generally high in that they are developed by test experts, pilot tested, and undergo revision prior to publication and use; and c) administration procedures are standardized and the test items are designed to rank examinees for the purpose of placing them in specific programs or instructional groups. Stewart and Kaminski report that local norms, based on the test performance of students from a specific locale, have the added advantage of providing meaningful information regarding average performance, for example, in a particular school or school district. These authors report many other advantages of using local norms, including that they decrease the likelihood of bias in educational decision-making because a student's test performance is compared to other students whose demographic and background factors are similar. In addition, they afford school systems the opportunity to compare data on students' educational outcomes to instructional curricula to which students have already been exposed. Furthermore, local norms are useful in facilitating decisions such as identifying the educational needs of students, determining standards for student progress, and identifying and making decisions about students' eligibility for Chapter I, English as a Second Language, and academically gifted programs. Finally, these norms are useful for identifying students at risk for school failure.

The predominant criticism of NRTs is that their content is seldom aligned with curricular content taught in educational settings (with the exception of locally normed tests). Good and Salvia refer to the match between the items on a norm-referenced achievement test and the content taught in a curriculum as content validity. The underlying assumption of content validity is that the items on a NRT should correspond to the content of the curriculum taught in a classroom. Results of a NRT devoid of content validity make it difficult to determine effective interventions that are needed for a student experiencing academic and/or behavioral challenges. Also, NRTs do not allow for monitoring academic progress over an extended period of time; instead, they provide an index of achievement or performance in comparison to a norm group at one specific point in time. Furthermore, an underlying assumption of NRTs is that examinees have had opportunities to acquire skills and experiences comparable to those of examinees in the norm group. If disparities exist between examinees and the norm group in terms of skills and experiences, the conclusions based on the examinee's test performance may be misleading. “When a child’s general background experiences differ from those of the children on whom a test was standardized, then the use of the norms of that test as an index for evaluating that child’s current performance or for predicting future performances may be inappropriate” (Salvia & Ysseldyke, 1991, p. 18). This potential problem is tied to the issue of cultural fairness, which has been the subject of significant consideration in test development and research. Essentially, no test is completely culturally fair. The responsibility for school practitioners is to ensure that a child's level of acculturation (i.e., the extent to which an individual has adjusted to the culture in which he or she lives) is considered when choosing a NRT. Flanagan & Ortiz have done extensive research on acculturation and language differences in the assessment of diverse children, and they provide in-depth information and guidelines on this topic.

With respect to limitations of local norms, Stewart and Kaminski cite misinterpretation as a primary disadvantage. Like the point made by Salvia and Ysseldyke, the group to whom a child's performance is compared should be well-defined (e.g., the age and grade of the students comprising the reference group as well as the size and stability of the group). Also, information about the measures administered to students as well as how the scores were derived should be provided. Stewart and Kaminski emphasize that the knowledge of how a student's performance in a particular subject area using local norms compares with that of their performance using national norms is quite significant. For example, a child may perform in the High Average tier of the local norm group in a particular subject area yet perform in the Below Average tier when tested on a nationally normed test in the same
subject area. To report how the child performed only in relation to the local norms would be misleading.

THE STATUS OF NORM-REFERENCED TESTING

In part, due to the shortcomings of NRTs, other types of assessments are used to assess individuals' aptitudes and abilities. Such methods include Curriculum Based Measurement (CBM) as articulated by Shinn; the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) as described by Good, Gruba, and Kaminski; and Curriculum Based Assessment (CBA) as described by Gravois and Gickling. Collectively, these methods use curriculum materials as the basis for assessing and monitoring students' academic progress. CBM and DIBELS assessment results are linked directly to instructional interventions, whereas CBA is used primarily to assess and modify a student's instructional environment for the purpose of placing the student in the most appropriate curriculum.

The Individuals with Disabilities Education Improvement Act passed in 2004 allowed states, for the first time, to use a student's response to scientific, research-based interventions as a basis for determining eligibility in the Specific Learning Disability (SLD) category of special education programs. This procedure, commonly referred to as Response to Intervention (RTI), is a significant shift from the traditional ability-achievement discrepancy model used in determining SLD eligibility. Although states have the option to continue using the ability-achievement discrepancy model, RTI, undoubtedly, will decrease the role of NRTs in qualifying students as learning disabled. Be that as it may, it is indisputable that NRTs facilitate meaningful comparisons between a student's test performance and that of the student's same-age peers in a test's norm group, and they will continue to be used and valued in the general education and special education arenas. It is also evident that CBA, DIBELS, CBM, and RTI procedures play a significant role in instructional decision making and in the early 2000s were gaining an increasing role in special education eligibility decisions. Using a combination of these two assessment paradigms may well be the optimal solution for serving the educational needs of all children.

See also: Standardized Testing

BIBLIOGRAPHY


*Copyright 2003-2009 The Gale Group, Inc. All rights reserved.*