Kingston, N. (2008). Norm-referenced tests. In N. Salkind (Ed.), Encyclopedia of educational psychology. (pp. 735-739). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412963848.n197

Norm-Referenced Tests

A norm-referenced test is one that is designed to facilitate interpretations of scores by comparing scores based on the ordering of examinees within a well-defined group of interest. Two important definitions associated with norm-referenced interpretations are *percentiles* and *percentile ranks.* A percentile is a test score below which falls a certain percentage of the scores. A percentile rank is the percentage of people who have a score lower than the score of interest.

Since about 1905, norm-referenced interpretations have been a dominant approach to making test scores meaningful to both educators and the public, although criterion-referenced interpretations have been gaining in prominence over the past several decades. A special type of criterion-referenced interpretation, known as *standards-based interpretation*, has been gaining in popularity since No Child Left Behind was enacted into federal law in 2002.

Score Interpretation

By itself, a raw score on a test has no meaning. Knowing an examinee answered 21 questions correctly on a math test is useless by itself. Faced with such useless information, one might ask any of a number of questions.

- How many items were on the test?
- What content within math was covered?
- What related content was excluded?
- What was the cognitive complexity of those items?
- What item formats were used (multiple-choice, problem solving, proofs)?
- How well did other examinees perform on this same test?
- What do we know about other achievements of examinees with similar scores?

Answers to each of these questions will raise other questions. Over time, the meaning of test scores accrues as users become familiar with the characteristics of those scores and the relationships those scores have with variables of interest.

Test makers try to facilitate this development of meaning by creating score scales that support the intended primary inferences. One such approach is referred to as *norm-referenced*—the comparison of the performance of an examinee with the performance of other examinees in a meaningfully defined group. Such interpretations may be particularly useful when determining how to allocate insufficient resources, such as if there are more applicants for an educational

program, school, university, or job than there are openings. For example, norm-referenced tests might be used as a significant piece of information in determining which students should be placed in a remedial or gifted program.

If resource allocation decisions were simple and there were 12 spots in a remedial program, one could simply admit the 12 students with the lowest scores. But most real-world resource allocation problems are more complex and somewhat elastic. Thus, developing expectations over time (and thus sometimes admitting more or fewer students into such a program) is facilitated by normative data. Therefore, policymakers might often prefer for a program to be made available for any students in the bottom 10%, rather than for a fixed number of students.

Normative expectations can also serve to facilitate group comparisons, for example, whether or not students in a school or district are performing as a group better than those in other schools or districts. Whether or not such differences are interpreted correctly, they can influence the perceived desirability of neighborhoods and, thus, real estate prices.

Historical Roots

The use of normative interpretation for test data has its roots in the work of early psychologists such as Wilhelm Wundt and Francis Galton in the late 1800s. These psychologists looked at the distributions of various measures and noted the typical normal distributions.

In 1905, French educator Alfred Binet invented the intelligence test. Scores were expressed as a mental age that could be compared with a student's chronological age to help make educational placement decisions. In 1916, U.S. psychologist Lewis Terman developed a revised version of Binet's test, the Stanford–Binet, and changed the score-reporting scale to the ratio of mental age to chronological age. The resulting IQ scores could be compared regardless of the age of the students, but proved to have more high and low scores than the normal distribution predicted. For this and other reasons, the ratio IQ score was replaced by a deviation IQ score with a mean of 100 and a standard deviation of either 15 (Wechsler) or 16 (Stanford–Binet). Deviation IQ scores were supplemented with information regarding the percentage of people with lower scores. In 1963, Robert Ebel coined the terms *norm-referenced* and *criterion-referenced* tests.

Normative Approaches

A common approach to providing normative information is the use of grade- or age-equivalent scores, just as Binet used more than 100 years ago. The general public finds this approach intuitively appealing because it attaches test performance to a concept with which they are very familiar—grade (or age) level. Student performance is indicated as being at the grade and month where the average child receives that score. Operationally, this is done by creating a concordance table, listing every possible score and the corresponding grade and month for which that score is the average score. Following is a raw score—grade-equivalent score concordance table for a hypothetical fourth-grade reading test with 15 possible raw scores (0–14).

There are two issues in particular that influence the interpretability of grade- and age-equivalent scores. First, such scales are not interval level. That is, the difference between a reading score of 1.1 (a student in the first month of the first grade) and 2.1 represents a greater difference in achievement than the difference between a 7.1 and an 8.1. Thus, it can be misleading to report growth scores or averages.

Second, consider two students who score a 6.5 grade-equivalent score on a science test. One student is in the second month of third grade and the other is in the second month of eighth grade. These students have been exposed to different parts of the science curriculum. It is likely that the younger student answered correctly almost all questions about the parts of the curriculum to which she or he had been exposed, but did not do as well on questions from those areas of the curriculum not yet studied. On the other hand, the older student may well have answered a small portion of the items correctly for all topics on the test. Furthermore, we would expect that 3 years from now, the younger student will surpass the science achievement of the older student.

An alternative approach to providing normative information is to provide percentile ranks for each raw (or scaled) score. This focuses comparisons within a grade (or age) level and thus avoids the second issue mentioned for grade-equivalent scores. The first issue (non-interval scale measurement) is also true for percentile ranks. On the other hand, the general public is not as familiar with percentile ranks as it is with grade level or age. An additional issue with percentile ranks is that they make it difficult to show

Table 1 Hypothetical Raw Score–Grade-Equivalent Score Concordance Table

Raw Score	*Grade-Equivalent Score*	*Raw Score*	*Grade-Equivalent Score*	*Raw Score*	*Grade-Equivalent Score*
0	2.2	5	3.9	10	4.7
1	2.6	6	4.0	11	4.9
2	2.9	7	4.2	12	5.2
3	3.2	8	4.4	13	5.4
4	3.6	9	4.5	14	5.7

growth. If students whose true achievement is at the 70th percentile in the fall make typical progress during the school year, they will have higher raw scores in the spring, but they will remain at the 70th percentile. Thus, within-year (or across-year) typical growth is reflected by no change in percentile ranks. Smaller-than-typical growth will be reflected by a drop in percentile rank.

Norming Process

The development of stable, accurate norms is a multifaceted, complex, and logistically difficult process. First, one must identify the normative population. Is it *all* ninth-grade students, just public school students, or public school students for whom English is their primary language of instruction? Should special education students be included? What about home-schooled children? At its best, the normative population should reflect the group within which students will be compared. However, sometimes there are multiple comparison groups, and so multiple norms can be desirable.

Once the normative population is defined, a sampling plan must be developed. Some tests (such as the SAT or ACT) provide normative information based on naturally occurring examinees. Such norms reflect the user population, but are inappropriate or misleading if one wanted to make inferences that went beyond that population, such as all college-bound seniors rather than college-bound seniors who took the SAT (or ACT).

Often, it is considered desirable to define the population of interest and, in a special study, collect data from a national probability sample (a sample that contains randomly selected examinees). Improved norms estimates (greater precision) can be derived from stratified random sampling; that is, breaking down the sample into subsamples. Stratified random sampling increases the precision of norms estimates most when the stratification variable is correlated with test scores.

Normative data must be gathered at a point in time for which the norms are most appropriate, but different test users will choose to use the test at different times. It is too difficult and expensive to collect normative data for each month or week of the year, so norms are typically collected at two different times for each grade (or age), and results are interpolated or extrapolated for other months.

Once raw data are received, those data must be adjusted. Data from each stratum are weighted to account for differences between the actual proportions collected in each stratum and the true proportions of that stratum in the population. For example, for the norming of an achievement test series, if a researcher stratified based on public and private schools, and after collecting the data, 70% came from public school and 30% from private schools, those data must be weighted to reflect that, in actuality, 86% of students attend public school and only 14% attend private schools. Thus, each public school student in the sample would get added into the distribution as 1.23 students (86/70), and each private school student would get counted as .47 students (14/30).

Even though an overall sample used in norming may be very large, at any particular part of the score scale there might not be very many examinees, and thus the data distribution might be jagged even though the underlying variable is distributed more smoothly. There might be scores that no one in the norming sample obtains, although when the test is administered operationally to a larger group, those scores would be

obtained. Thus, the data collected in norming studies are often smoothed using any of a variety of statistical techniques. It is at this stage that data are interpolated or extrapolated for months from which no or insufficient data were collected.

Test Design and Development Issues

Content Coverage

Norm-referenced tests are used by school districts throughout the country, and although those districts have overlap in their curricula, they also have significant differences. Thus, content appropriate within one grade in a certain district might be most appropriate at a higher or lower grade in another district. Also, many of the decisions based on norm-referenced tests are for very high- or very low-achieving examinees, and such tests often have both material from earlier stages in the curriculum (for example, fourth-grade material on a sixth-grade test) and later in the curriculum (such as eighth-grade material on that same sixth-grade test). For both of these reasons, norm-referenced tests tend to have broader content coverage than comparable criterion-referenced tests.

Item Difficulty

A test will have the greatest accuracy of measurement for the greatest number of examinees if all test items are of middle difficulty. Middle difficulty is achieved when half of the examinees know the answer to a question. On tests where guessing can be a factor, this means that somewhat more than 50% of the examinees will answer the questions correctly. For a test consisting of five-choice multiple-choice questions, this means that 50% will answer correctly because of their knowledge and 10% (one-fifth of the remaining 50%) will answer correctly by guessing, for a total of 60% correct. Thus, if maximizing the average score accuracy was the primary concern, there would be no need for very easy or very difficult items. However, as stated before, many of the decisions based on norm-referenced tests are aimed at students at the high and low ends of the achievement continuum, and thus there is a need for accurate measurement at the ends of the score scale, resulting in easy and hard items in addition to middle difficulty items. A balance between these conflicting item difficulty requirements must be struck.

Special Topics

Issues Related to Systems of Norms for Vertically Scaled Tests

When there are multiple test forms for different grade or age levels, there are additional issues associated with developing test score norms. For example, if norms are developed separately for the third-grade form of a test and the fourth-grade form, it is desirable that regardless of the form of a test a student takes, students with the same scaled score and grade level have the same percentile rank.

Participation Rates

It is getting increasingly difficult to get a random sample of schools to agree to participate in a norming study. Many schools feel that their students are overtested. This is particularly true at the high school level. For example, in the 2005 science administration of the National Assessment of Educational Progress (NAEP), 87% of invited 4th-grade schools participated, but only 83% of 8th-grade schools and 76% of 12th-grade schools. At all three grade levels, more small schools chose not to participate. Additional schools were invited to replace those that did not participate, but it is likely that the resulting sample was unrepresentative of the nation as a whole.

Most importantly, NAEP participation rates seem to be much better than those obtained by commercial test publishers. Thus, there are many questions regarding the representativeness of published norms.

Motivation

There is concern (and some evidence) regarding the motivation of students participating in norming studies. If students (or their teachers) are not as motivated during a norming study as they are during an operational test administration, then it will appear as if the test is harder than it actually is, and norms will overestimate student relative performance.

Political Controversy

Norm-referenced and criterion-referenced tests have been caught in the crossfire of political debate on the quality of American public education. One faction feels that it is important to provide national norms to allow schools to demonstrate whether they are

performing adequately compared to schools throughout the nation. Another faction feels that norm-referenced interpretations are flawed in that if all schools improve, half of all students (by definition) will still be below average.

Percentile Data

Percentiles are ordinal-level data (for example, the difference in achievement between the 98th and 99th percentile is much larger than the difference between the 50th and 51st percentile). Thus, it may be misleading to take an average of percentiles. Instead, one should average the scaled scores associated with those percentiles and take the percentile of the average score to represent the normative performance of a group.

Individual Versus Group Norms

The average scores of groups do not vary nearly as much as the individual scores of examinees within the population. Thus, normative information to illuminate group (for example, school) performance should be based on the distribution of group averages and not on the percentile rank of the average examinee in that group. For example, a high-performing school might have an average score of 310 on some hypothetical test. A score of 310 might have a percentile rank of 75 when applied to an individual, but an average score of 310 might be better than 98% of all schools.

Adoption of Norms for Use in Customized State Tests

Standards-based tests such as those required by No Child Left Behind require a close match to state-specific curriculum frameworks. Norm-referenced tests are usually developed to be broader (and correspondingly less deep) than standards-based tests. Many states would like to provide national normative interpretations for their state standards-based test. But norms are developed for a specific test and a specific population. Some states and publishers have tried to augment norm-referenced tests with items necessary to provide the depth required of a state standards-based test, equate the state-augmented test to the nationally normed test, and use the national norms to estimate how students nationwide would have done on the state-specific test. To the extent that the curriculum frameworks differ from state to state, the results from such a process might be significantly different from those obtained if a norming study were done administering that state's test throughout the nation.

Neal Kingston

See also Criterion-Referenced Testing; Grade-Equivalent Scores; Measurement; Standardized Tests

Further Readings

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement.* Westport, CT: Praeger.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1988). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement.* New York: American Council on Education.