Chapter 6 Educational Extension

Interpreting Standardized Test Scores

Standardized Tests

Ebel and Frisbie (1991) define a standardized test as one that has been constructed by experts with explicit instructions for administration, standard scoring procedures, and tables of norms for interpretation. Standardized tests measure individual performance on a group-administered and group-normed test. Standardization means that the examinees attempt the same questions under the same conditions with the same directions, the time limit for taking the test is the same for everyone, and the results are scored with the same detailed procedure.

The preparation, editing, and analysis of standardized tests are undertaken by experts in the field covered. First, these experts write a battery of content questions related to the subject field that should be answerable by the average, well-informed student at the targeted grade level. These questions are then tried out on a representative sample of students in the subject area at the specified grade level from all kinds of schools in all parts of the country. Based on the results obtained from the representative sample, the test is revised, and a final version of the exam is written. The exam is then administered to a sample of students that is larger and more carefully selected to represent the target subject area and grade level. This latter sample of students forms the norming group against which all subsequent student scores will be compared. The final step in the development of the standardized test is the production of a test manual that provides clear directions for administration and scoring of the test, as well as certain information related to the test characteristics and interpretation of scores.

Standardized tests are particularly useful in comparing the performance of one individual with another, of an individual against a group, or of one group with another group. Such comparisons are possible because of the availability of the norming-group data and the uniformity of the procedures for administration and scoring the test. Thus, you could compare a single school with others schools in a district, compare districts, compare states, or compare students in one state with all students in the nation. For example, suppose a student scored 80% on a reading test. For most purposes, this is sufficient information for evaluation purposes, but at times you might want to know how a reading score of 80% compares to the reading scores of other students in the district, state, or nation.

Standardized test results generally include a percentile norm, an age norm, and a grade-level norm, or a combination of these norms. Teachers are often called upon to interpret these norms for parents. Therefore, you need a basic understanding of their meaning. At the most basic level, a percentile rank of 85 indicates that 85% of the norm group performed more poorly and 15% performed better than the individual in question. An age rank of 15.6 means that, regardless of chronological age, the student got the same number of right answers as did the average 15½-year-old in the norm group: The number to the left of the decimal point represents years, and the number to the right of the decimal point represents months. Grade-level norms are the most widely reported and the least useful of the measures. Grade-level norms are reported as grade-equivalent scores, with the tenths place representing the months in a 10-month school year. The only real value of grade-equivalent scores is to determine if the test was too easy or too difficult for students scoring at a particular level.

Although these reasons present a strong case for the use of standardized tests in educational decision making, it is important to recognize that a too narrow or too broad interpretation of results can lead to oversimplifying a complex situation. For example, one
English class performs more poorly on grammar than another. Can you assume that the teacher is doing a poor job of teaching grammar concepts? Probably not, and certainly not without more information about the composition of the two classes. The poor performance could be explained by a large number of mainstreamed special-needs students in the first class or by the fact that all school assemblies and announcements occur during second period, reducing instructional time during that period.

In your future classroom you will analyze standardized test data to develop action plans to help struggling students. To correctly interpret these data you need to gain background knowledge about the normal distribution curve and standard scores that are commonly used in statistical analysis.

The characteristics of the normal curve make it especially useful in education and in the physical and social sciences. In this overview, we will introduce you to the basic concepts associated with the normal curve and the calculation of various standard scores.

**Normal Curve**

The normal distribution is a continuous probability distribution that describes data that cluster around a mean or average. The graph of data is bell-shaped with a peak at the mean (see Figure 1). Most evaluators concerned with students’ standing within a group make use of the normal curve. This curve is commonly called the natural curve or chance curve because it reflects the natural distribution of all sorts of things in nature. Such a curve is appropriately used when the group being studied is large and diversified. The curve makes it a lot easier when interpreting and communicating student data. Within a classroom, the curve can be used to give teachers an idea of how well a student has performed in comparison with classmates.

An understanding of the normal curve requires a basic knowledge of the concept of variability; that is, you must understand standard deviation. The standard deviation is a measure of the extent to which scores are spread out around the mean. The greater the variability of scores around the mean, the larger the standard deviation. A small standard deviation means that most of the scores are packed in close to the mean. A large standard deviation means that the scores are spread out. When all the scores are identical, the standard deviation is zero.

The normal curve is a mathematical construct divided into equal segments. The vertical lines through the center of the curve (the mean) represent the average of a whole population on some attribute. For example, in a large population, the average (mean) IQ would be 100. Because this number represents the mean, it would occur most often and appear at the highest point on the curve. To the left of the highest point, each mark represents 1 standard deviation below the average. To the right of the highest point, each mark represents 1 standard deviation above the average. Neither side of the curve would touch the baseline, thus showing that some extreme IQ scores might exist. As shown in Figure 1, about 34% of the named population on an attribute will be within 1 standard deviation below the mean, and about 34% of that population will be above the mean. About 13.5% of the identified population will be in the second deviation below the mean, and 13.5% of the population will be in the second deviation above the mean. About 2% of the population will fall in the third deviation below the mean, and an equal portion will fall in the third deviation above the mean. Finally, about 0.1% of the population will fall in the extreme fourth deviation below the mean, and 0.1% will be in the fourth deviation above the mean. Some of the many things that are subject to the normal distribution are weights and heights of animals, average temperatures over an extended period of time, and margins of error on measurements.
Indeed, most schools use normal curve and standard scores when reporting the results of standardized tests.

**Standard Scores and Percentile**

There is an old saying that “you can’t compare apples and oranges.” However, this is not always the case. Any interval or ratio variable can be converted to a standard unit of measurement called standard score. They can then be compared.

Most schools report student performances in terms of standard scores such as $z$ scores, $T$ scores, and stanine scores, as well as in terms of percentile. These methods use the normal distribution curve to show how student performances compared with the distribution of scores above and below the mean. Standard scores provide a standard scale by which scores on different evaluative instruments by different groups may be compared reasonably (see Figure 2). The various standard scores and percentile say the same thing in slightly different ways.

Note that $z$ scores and $T$ scores correspond to the standard deviation of the population scores: They tell us how far above or below the mean in standard deviation units raw scores lie. Both $z$ scores and $T$ scores indicate the number of standard deviations that a particular raw score is above or below the mean of the raw score distribution. Thus, a score falling 1 standard deviation below the mean would have a $z$ score of –1, a score falling two standard deviations above the mean would have a $z$ score of +2, and so on. The use of negative numbers is avoided by converting $z$ scores to $T$ scores. This is done by multiplying the $z$ score times 10 and adding a constant of 50.

$$T = 10z + 50.$$ 

Many schools use stanine (standard nine) when reporting student performance, with a stanine of 1 representing the lowest performance and a stanine of 9 the highest. These nine numbers are the only possible stanine scores a student can receive. Stanines use the normal distribution in grouping scores into nine categories, with a mean of 5 and a standard deviation of 2. Figure 3 gives the stanine score distribution and the percentage of scores that will fall into each category.

Another type of score is the percentile. Percentile scores are often confused with percentage correct. However, the percentile score indicates the percentage of the population whose scores fall at or below that score. A score of 20, for example, would have 20% of the group falling at or below the score and 80% above the score. Of course, the 50th percentile would be the mean (see Figure 2). Equal differences between percentile scores do not necessarily indicate equal differences in achievement.

Converting raw scores to standard scores allows you to compare scores on different assessment instruments. You could compare a student’s performance in mathematics with science or reading with English or any other tested areas. For example, if a student obtains a score of 65 on his or her math midterm and 70 on his or her history midterm, on which test did he or she do better compared to the rest of the class? If converting the student’s scores to standard score yields a $z$ score of 1.5 in math ($T$ score of 65) and a $z$ score of 1.0 in history ($T$ score of 60), you can conclude that the student did better in math.

**POINTS TO PONDER**

1. When teachers “grade on the curve,” they expect students’ scores to create a normal curve. Therefore, most students score a C and very few score As or Fs. How do you feel about this grading approach?
2. Central measures of tendency (mean, mode, median) are statistical measures used to measure student performance. In a perfect normal curve, all three statistics align in the center of the curve. Remove the face cards and jokers from a deck of playing cards. Deal out two cards at a time until you’ve generated 100 two-digit numbers.
   a. Calculate the mean (average) by adding the numbers and dividing by 100.
   b. To determine the median (or the central number), arrange the numbers from lowest to highest. Cross off the numbers at the outer edges, then the next pair and so forth until you reach the middle. The average of the last two numbers is the median.
   c. The mode is the number that appears the most frequently.

3. Arrange to review standardized test results from a local school district. Be sure to blacken out any student demographic information to protect the students’ identity. Which statistical measures are used to analyze the data? If these were your students, what could you do to address the areas of perceived need?

Reference
