Scaled Scores and Performance Levels: The Nuts and Bolts behind Them and Issues with Their Use

Stuart Kahl
RMC Research Corporation
May 2019

When state testing officials are asked what they wish educators and parents understood better, one of the answers that arises is scaled scores and performance levels – the most common ways student test performance is reported these days. What follows is a "middle-level" explanation of these, which may be of use in helping various consumers of test results better understand the information they are provided by testing agencies and companies. ("Middle-level" means somewhere between simplistic and sophisticated in terms of measurement content.)

At the risk of insulting the intelligence of readers, I'm going to start by mentioning some basic measurement concepts to which all early elementary students are exposed. (Think of this as a reminder of fun childhood activities, but very relevant.) It's not unusual for teachers to ask their young students to measure the length of a desk using different sized paper clips or to use shoes to measure the length of a room. Of course, coming up with different numbers for the length of the same object, the kids learn of the need for standard units such as inches or feet. These allow appropriate comparisons of lengths of different objects. The kids also learn something about frequency distributions when they build bar graphs by pasting paper squares in columns above the names of different objects they've counted.

Frequency Distributions of Scores

Now to move on from these fond memories, let's think about test scores listed along the horizontal axis of a graph. Above each, we can place a vertical bar the height of which corresponds to the frequency (number of occurrences) of the score on the vertical axis. If there are a wide range of scores and a large number of students and we connect the tops of the bars, typically we'll get something resembling a normal or "bell" curve. By the way, contrary to popular belief, test makers didn't invent the normal curve. There are lots of things that are naturally normally distributed or close to it. Take men's weights for example. How many grown men weigh 40 pounds? Not many. And how many weigh 300 pounds? Not many. But as we move from the extremes toward 100 or 150 pounds, there are more and more men at the more typical weights. Statisticians have come up with mathematical formulas for curves that fit various distributions of "real things" or phenomena. And those formulas enable statisticians and psychometricians to do some really important stuff.

Test scores can be normally distributed or close to it. Of course, an especially easy test may lead to a curve with the hump shifted more to the right toward higher scores, and an especially hard test may lead to a curve with the hump shifted more to the left toward lower scores. But for our purposes, close to normal is good enough. The **mean** and **standard deviation** of a set of scores are especially important to us. The mean, of course, is the average score. In a perfectly normal distribution, it corresponds to the score on the horizontal axis directly below the highest point of the curve.

The standard deviation is a measure of the spread of the distribution. A flatter distribution has a larger standard deviation than a tall narrow distribution, both using the same scale on the horizontal axis. A particular score's deviation from the mean is simply that score minus the mean score. One might think that the average deviation of scores from the mean would be a good measure of spread or variability of scores. However, the scores below the mean have negative deviations that cancel out the positive deviations of scores above the mean when all are added together; and averaging those always yields zero. However, squaring the deviations from the mean yields all positive numbers, so the average of the squared deviations makes a good measure of variability. With a slight fudge factor – dividing the sum of squared deviations from the mean by the number of scores minus one instead of by the number of scores – we get the variance of the distribution. (For large n's, dividing a number by n-1 instead of by n makes very little difference.) The square root of the variance is the standard deviation.

Standard Scores and Scaled Scores

Let's get back to "real" test scores. Suppose Jamie took Test X and Test Y and scored 75 and 78 on them respectively. On which did Jamie do better? Answers like "Test Y" or "too close to really say" might seem reasonable. Now let me give you some information on the means, $\bar{x}$ and $\bar{y}$, and the standard deviations, s_x and s_y, of the two score distributions:

Test X: $\bar{x} = 50$ and $s_x = 8$

Test Y: $\bar{y} = 80$ and $s_y = 10$

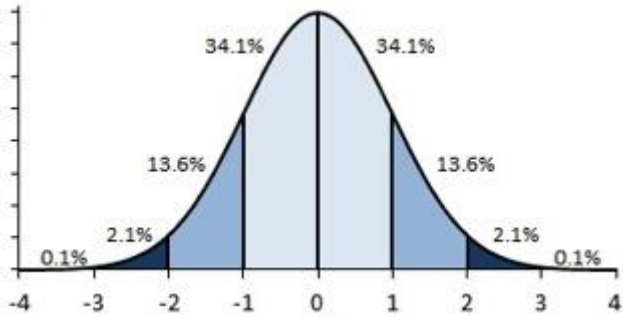
Clearly with this information, we should conclude that Jamie knocked the socks off Test X, but performed a little below average on test Y. To make an appropriate comparison of the two scores, we should transform them to a common scale. We can transform them each to a z-score, which is the number of standard deviations away from the mean a student scores. Z-scores are a form of standard scores.

Back to Jamie's tests, the score on Test X of 75 is 25 points above the mean of 50. How many standard deviations are in 25 points? 25/8 = 3.125, so $z_x = +3.125$. The score of 78 on Test Y is 2 points below the mean of 80. That is 2/10 of a standard deviation, which means $z_y = -0.2$. The z-scores on the two tests are appropriate scores to compare because the non-comparable raw scores were transformed to scores on a common or standard scale – the z-score scale. Again,

relative to the performance of others, Jamie did much better on Test X. Of course, we should know a lot more non-quantitative information about the two tests. If they were both math tests, then Jamie's results would be quite unusual, since we would anticipate more consistent performance. If Test X and Test Y were successive unit tests in math and Jamie was hospitalized when the second unit of instruction was delivered and couldn't study or do homework, then that might explain Jamie's considerably worse performance on Test Y. If the tests were in two different subjects and Jamie was a math wizard, but typically a below average performer in social studies, then that could explain the different results on a math test, Test X, and a social studies test, Test Y.

If you want a formula for a z-score, just think of the definition – the number of standard deviations away from the mean. Compute the distance from the mean, then find the number of standard deviations in that distance: $z_x = (x-\bar{x})/s_x$. If we know the z-scores corresponding to a set of raw scores on a test, we can transform those z-scores to any scale we want. Suppose we wanted to transform the z-scores on a third-grade standardized test to a scale with a mean of 350 and a standard deviation of 25. As an example, a z-score of +0.4 is 4/10 of a standard deviation above the mean. That would give us a score of 350 + .4(25) = 360 on the new scale. If you want a formula for a score x on a new scale, given z-scores, just solve the formula for a z-score for x. You get $x = \bar{x} + z_x s_x$.

A picture is worth a thousand words. Much of what is discussed above regarding Tests X and Y is shown below.



Test X	18	26	34	42	50	58	66	74	82
J's Test X raw score								75	
Test Y	40	50	60	70	80	90	100	110	120
J's Test Y raw score					78				
IQ	40	55	70	85	100	115	130	145	160
New Gr 3 scale		275	300	325	350	375	400	425	450

The graphic shows a normal curve, and the horizontal scale immediately below it is in standard deviation units. Those values are z-scores. The percentages are the percentages of the total area under the curve within different regions. Thus, 68.2 percent of the area, which could correspond to the scores of 68 percent of students, is within one standard deviation from the mean in a

normal distribution. Looking back at the transformations we made earlier between different scales, we see that these are linear transformations. The formulas are equations of lines – remember y=mx+b. Going from z-scores to another scale like those shown below the graphic just shifts the mean from zero to something else and stretches (or compresses) the distribution. However, to show the stretched distributions on a scale counting by one, even newsprint wouldn't be large enough to depict the scores shown. So instead, we accomplished the same thing by re-labeling the horizontal axis.

People sometimes mistakenly believe that converting raw scores to z-scores makes a non-normal score distribution normal. That is not true. Linear transformations of scores do not change the shape of a distribution. Furthermore, the relative distances between different scores are preserved. I mentioned before the distributions of especially easy or hard tests, which would not be normal. For these, what's important is that the mean and standard deviation of the z-scores would still be zero and one respectively. It's just that the mean of zero would no longer correspond to the highest point of the skewed distributions.

Scaling and Equating

Below the curve on the previous page, the Test X and Test Y scores (including Jamie's scores) are raw scores. The others (IQ and New Gr.3) are **scaled scores**. To make an appropriate comparison between the two test results, we "scaled" the scores of the two tests by converting them to z-scores; and we know we could then transform the scores again to any scale we wanted to use. By the way, Jamie's reported "raw" scores could have been percents of total possible points Jamie earned instead of points earned. However, the score distributions would still have different means and standard deviations, and thus reflect two different scales, and we would still have to transform scores to standard scores to make appropriate comparisons.

Now that we understand z-scores, I can tell you that state and commercial testing programs do not use the z-score approach for their initial transformation of raw scores. They use something called Item Response Theory (IRT) for that step, but the result is still a scale with a mean of zero and standard deviation of one. I could explain the more sophisticated IRT approach in four or five pages, but for our purposes here that's unnecessary. The point is that the initial scaling step followed in a new testing program transforms the raw scores to the scale which is easily transformed again to whatever scale is desired. (The new grade 3 scale mentioned earlier exemplifies this.) But guess what….we're not home free yet.

Suppose there are different forms of a test, both intended to measure the same thing. They could be two forms of a state's math assessment used in the same year or different math forms used in the same grade in different school years. We can make sure that the test development specifications call for different test forms to include similar numbers of items in different content

categories (e.g., algebra and geometry in mathematics) or similar numbers of items of different types, and we can make the two test forms roughly equivalent in difficulty. However, to be sure that scores across forms or across years are comparable – e.g., that a score of 340 has the same meaning relative to students' capabilities regardless of the form or year – the forms have to be **equated** in terms of difficulty. Years ago, score adjustments on forms may have been made by adding or subtracting points to kids' scores. IRT techniques actually accomplish equating more accurately.

So scaling is not enough. We could convert raw scores on different forms to the same scale (e.g., a mean of 350 and standard deviation of 25), but not only would that fail to take into account different difficulties in forms, it would also not allow the accurate measurement of different types of academic "growth." For pre-post testing using commercial tests at the beginning and end of a school year, the mean score should definitely not be the same for both administrations. On state tests, if performance of third graders is improving in successive years due to instructional improvements or teachers' greater familiarity with curriculum standards or if a group of students passing through a tested grade in a school is academically stronger than the group passing through the grade the previous year, then the mean scores should reflect the superior performance the second year. The tests have to be equated to allow that to be done accurately.

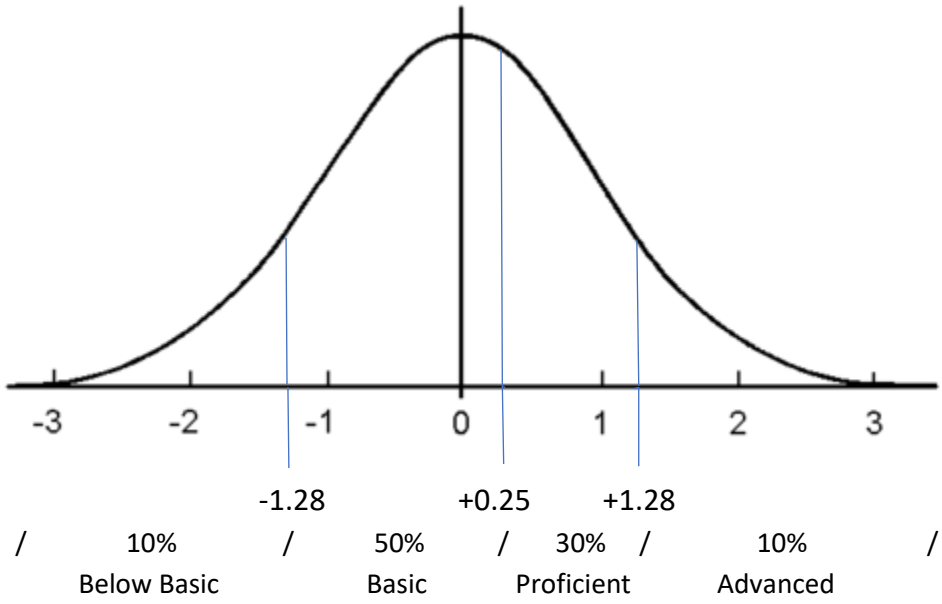
Performance-Level Reporting

Consistent with the requirements of IASA, NCLB, and most recently ESSA, state testing programs (and most commercial test series) do performance level reporting. That is, they have established **cut scores** on their scaled score continua that separate students, based on their scaled scores, into different categories, called performance or achievement levels. Many states use four levels, but they often differ in the names they give the levels. It is typical that the second from the top is called "Proficient" because of the federal law's emphasis on every school's percent of students proficient or above. For our purposes here, let's call the levels Below Basic, Basic, Proficient, and Advanced.

The cut scores separating the score ranges for the different levels are determined by a process called **standard setting**. An early step in this process is the development of performance level definitions. These are brief descriptions of what students at the different levels know and can do relative to the curriculum standards covered by the tests (and hopefully, by instruction). To oversimplify, the second level (e.g., basic) descriptor typically says the students at that level have a lot of basic knowledge and skills, but have a hard time applying them. Proficient students are said to be able to apply basic knowledge and skills to address relatively routine tasks or problems. Advanced students generally show deeper insight and are able to apply their knowledge and skills to more complex tasks or problems. The actual performance level

descriptors states use are longer than my examples, but they are still fairly general in nature. They are produced by content area specialists, but their acceptance is ultimately a policy decision as are curriculum standards and the final cut points on the assessments.

There are several different standard setting methods, but they generally involve panels of judges (educators and non-educators) who are asked to rate either student work or test items. After being familiarized with the test content and the performance level definitions, the judges are asked to match samples of student work or test items (based on item requirements) to the definitions. The aggregation of judges' ratings lead to cut scores on the test's raw score scale. Ultimately, approved cut scores are transformed to their corresponding scaled scores. The graphic below shows the results of hypothetical standard setting relative to our familiar distribution with a mean of zero and standard deviation of one. The process led to 10 percent of the students in the Below Basic level, 50 percent Basic, 30 percent Proficient, and 10 percent Advanced.



If the cut scores -1.28, +0.25, and +1.28 were to be transformed to that new distribution we mentioned earlier with a mean of 350 and standard deviation of 25, then those cut scores would be 318, 356, and 382. Assuming we're talking about statewide results in the first year of a testing program, with equating those cut scores would remain the same in subsequent years. However, improved performance would result in a distribution shifting to the right and perhaps with a slightly different shape. And there would be different percentages of students in the performance levels – most likely more students in the upper categories and fewer in the lower categories.

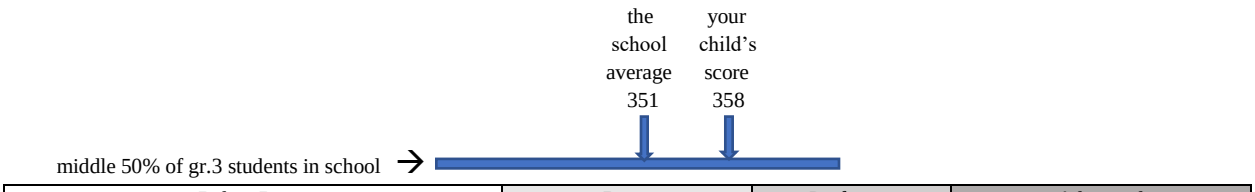
Many state assessment programs use "nice numbers" for their scaled-score cut points. For example, instead of choosing a particular mean and standard deviation for their score scale, officials could choose numbers like 340 for the Basic cut score (the minimum score for the Basic category) and 360 for the Proficient cut score. Knowing the cut scores on the IRT scale that emerged from standard setting and that two points determine a line, they would have two (x,y) ordered pairs defining the transformation line, an x being an IRT cut score and a y being the desired cut score on the new scale. Plugging those values into the equation of a line, y=mx+b, they'd obtain two simultaneous equations with two unknowns – the slope m and the y-intercept b. Solving for m and b, they would have the general equation they need to determine the new scaled score corresponding to any IRT score. The third cut score on the new scale almost certainly would not be a "nice number." C'est la vie. Nevertheless, as mentioned earlier, the order and relative spacing between student scores would be preserved when the scores are transformed from one scale to another – not unlike changing a temperature reading from Fahrenheit to Centigrade.

Issues with Test Report Statistics

Years ago, schools' mean scaled scores were what was emphasized in school reports and highlighted in the newspapers. Critics of such reporting felt that additional information regarding score distributions would be important. Performance level reporting does that. And for school accountability purposes as required by federal law, percent of students proficient or above is a reasonable metric. (The expectation of 100 percent proficient or above is another matter.) Also, performance level descriptors attach some underlying meaning to test scores in terms of students' capabilities. These general descriptors seem reasonable for describing typical students within performance levels, and cut scores provide useful targets for individual and group performance. However, at the individual student level, performance level reporting can be misinterpreted.

I often hear parents make claims such as, "My child scored Basic in math and Proficient in reading." So what does "Basic" mean in reading? Let's look at the very realistic distribution shown on the previous page to answer that question. The student could have scored anywhere between 318 and 356 to be designated as at the "Basic" level. Those scores are roughly a standard deviation and a half apart! In percentiles, were talking about somewhere between the 10th and 59th percentiles. (That's far apart, but be careful -- a percentile is just the percentage of student scores a particular score equals or exceeds. Percentiles cannot be subtracted or averaged.) The point is that a student whose score is just above the lower cut score for a performance level is in the same category as a person whose score is just below the cut score for the next higher level. They would be very different in terms of their capabilities. Does the same performance level descriptor really describe each performance or the two students' capabilities accurately? Also, the performance of a student scoring just below a cut score and that of a student scoring

just above the same cut score are virtually indistinguishable, yet those students are assigned to different performance levels with different descriptors. This is why a parent report from a state assessment program would usually display a student's scaled score in a graphic showing where the student performed within a level, as well as where he/she performed relative to other students in the school. The graphic below illustrates how such information could be displayed and is similar to parent report approaches some programs use.



School Percents of Gr. 3 Students:	7% Below Basic	47% Basic	34% Proficient	12% Advanced
State Percents of Gr. 3 Students:	10% Below Basic	50% Basic	30% Proficient	10% Advanced

A raw test score in isolation is meaningless. What does it mean to say that Joey earned 36 points on a test? A student's test score is only meaningful in a relative sense – relative to the performance of a referent group (e.g., students in the same grade in the school or statewide), relative to established standards of performance (such as the cut scores for performance levels), or relative to previous performance. The graphic above addresses two of these. When a testing program tests the same students in successive years, then performance relative to previous performance (growth) can be reported. There are several different ways assessment programs can represent academic growth. A discussion of these goes beyond the scope of this paper; and quite frankly, attaching meaning to extent of growth is more complex than simply comparing two numbers. How much academic growth during a school year is enough? What does a student's growth mean in terms of the student's capabilities? How much growth should be expected of a particular student in a year?

Occasionally, one might hear someone say something like, "Thank goodness performance levels have gotten us away from that darn normal curve." Well, nothing could be farther from the truth. The display on page 6 should make this obvious. Clearly, there is a continuum of student performance or capability underlying the performance levels and test scores along that continuum are almost certainly close to normally distributed. Yes, a bar graph with four bars of 10, 50, 30, and 10 units does not look much like a bell curve, but look at what's under the surface. In fact, it could be said that reducing a full score continuum to four levels represents a significant loss of information. The combined use of scaled scores and performance levels illustrated above is much more informative.

One final point about performance level reporting relates to competency-based education (CBE). There is a tendency among some CBE advocates to believe, or at least behave, as if students are either competent or not competent. This may have been a reasonable position forty years ago in the days of minimal-competency testing. Back then, competencies were defined very narrowly. For example, a math competency might have been "adding two three-digit numbers with regrouping." Expecting students to demonstrate competency with respect to that skill by answering three out of four almost identical multiple-choice questions correctly may have worked. But today's competencies are much broader and more complex, and therefore would be associated with underlying continua of performance quality. Testing in a CBE environment still needs to discriminate among a range of student abilities, and there is a need for enough measures to yield results that generalize beyond just a few contexts or problem situations. In CBE, a student may not be tested for record until he or she is ready, but there is still a range of performance among the "ready" students. And testers are still faced with the task of determining what performance along a continuum of performance quality is "good enough." "Competent" is a performance level. It is just as important for local educators as it is for state testing officials to recognize the importance of providing more information on a student's performance than that which is conveyed by a category name.