# Developing Criteria and Setting Cut Scores

Assessment data is fundamentally descriptive. To utilize it in evaluation, criteria must be established. A criterion serves as a benchmark or the standard for determining whether a performance or achievement meets acceptable levels.

Earlier we made the distinction between assessment and evaluation. Assessment data is inherently descriptive. For it to be used in an evaluation, we need to set criteria. The criteria we choose will form the basis for judging whether a performance or an achievement is acceptable.

Assessment criteria are not always absolute standards. For example, a standard may be universally applied, as in the case of visual acuity. A desirable vision acuity, regardless of your age or gender, is generally considered to be 20/20 vision. This is considered "normal" or optimal vision. It means you can see at 20 feet what a person with standard vision can see at 20 feet. This is the benchmark or criteria for what we call "perfect" vision, though some people have vision that's better than 20/20, such as 20/15, where they can see at 20 feet what a person with standard vision would need to be at 15 feet to see. Slightly worse vision, like 20/30, is still considered quite good and often doesn't require correction for most daily activities. In the United States, for instance, most states require a visual acuity of 20/40 or better in at least one eye, with or without corrective lenses, to pass the vision test for an unrestricted driver's license. This standard applies to all drivers regardless of their age. In this case, the pass score, the criteria, is 20/40 for everyone.

We might also use variable criteria to judge a performance or ability. For instance, the acceptable standard for academic performance in schools may vary depending on the grade level or developmental norms. A common criterion in education that varies depending on developmental stage is reading fluency, particularly words read correctly per minute (WCPM). This measure is used to assess a student's reading speed and accuracy, and the expectations change as a student progresses through grade levels. A part of an individual's natural development includes the notion that children naturally become more capable as they approach adulthood. Students in early elementary school may be expected to read at 30-80 WCPM, while high school students are expected to read at 150-200 WCPM, though the focus often shifts to comprehension and analysis at this stage

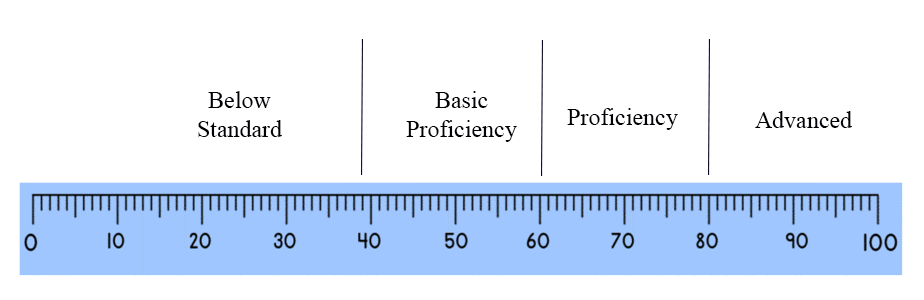
These ranges are approximate and can vary based on specific educational standards and assessment tools. The key point is that the expectation increases as students develop greater capacity. This criterion acknowledges that reading skills develop over time and that it's unrealistic to expect the same performance from a first-grade student as a high schooler. Using variable criteria allows educators to set appropriate goals for students based on their grade level, identify students who may be struggling compared to their peers, and track an individual's progress over time.

The point we are making here is that we need to set criteria in order to make criterion-referenced decisions. Unlike norm-referenced interpretations, we interpret these assessment results against the standards we set rather than as a comparison to others taking the test. The question you may be asking is how this is done.

#### Setting Cut Scores

While a criterion is a specific standard or benchmark that defines acceptable performance, a cut point (or cut score) is the actual score on the test that separates those who meet the criterion from those who do not. Determining cut scores involves a systematic process, often incorporating expert judgment, empirical data, and some politics.

While a single cut score may identify the passing score, most standardized criterion-referenced tests have two or three cut points that divide the test into distinct performance categories. Using three predetermined cut points on a test to classify student achievement, you might label the resulting performance categories as Below Standard, Basic Proficiency, Proficient, and Advanced. Decisions about a particular student's proficiency are determined by the specific category in which the score falls.



The process of setting cut scores, known as standard setting, involves careful consideration of test content, performance expectations, and potential consequences to ensure fair and meaningful categorization of student abilities.

There are two common approaches to setting pass criteria for standardized tests:

1.      **Norm-referenced method**: This approach sets the pass score based on the typical performance of test-takers in a norm group.

2.      **Criterion-referenced method**: This approach sets the standard that test-takers must meet, regardless of how others perform. Subject matter experts determine what knowledge or skills a minimally competent person should possess and set the pass score accordingly.

Each method has its advantages and is suitable for different purposes.

#### **Setting Cut points using a Norm Group**

A norm group is a representative sample of test-takers used to establish performance standards against which individual scores can be compared. Students in the norm group take the test under standardized conditions, and their collective performance creates a frequency distribution that represents the norm. This distribution represents the typical performance of student and can therefore be used to establish reasonable cut points.

Using a norm group to determine cut scores ensures that the performance standards are grounded in empirical data reflecting actual performance patterns within the population. This approach allows for meaningful comparisons and helps maintain fairness and consistency in interpreting test results, as every individual is evaluated against the same performance benchmarks established by the typical performance of student in the norm group.

In order for a norm to be valid it must be representative of the population, of adequate size, and up-to-date.

1.      **Representative Sample**: The group should accurately reflect the demographic and academic characteristics of the population for which the test is intended. It should include a wide range of abilities and backgrounds, including factors such as age, gender, socioeconomic status, ethnicity, geographic location, and educational background. It must have a proportional representation of all relevant subgroups within the target population, including an adequate representation of minority groups and potentially underrepresented populations.

2.      **Adequate Size**: The sample should be large enough to minimize sampling errors and provide stable norms.

3.      **Up-to-date**: A norm should be based on recent test administrations to ensure relevance, account for potential changes in population characteristics, and address the potential for inflated norms due to the Lake Wobegon effect.

**The Standard Normal Distribution.**The assumption of normality in educational achievement refers to the belief that most educational outcomes when measured across a large population, follow a normal or Gaussian distribution. This concept is fundamental in many statistical approaches used to interpret educational achievement and is crucial when determining cut scores.

The notion that cognitive and physical ability usually follows a normal curve distribution within a population is supported by various strands of empirical evidence and theoretical reasoning:

1. **Historical Data on Cognitive Abilities**: Research in psychometrics has consistently shown that scores on cognitive ability tests tend to follow a normal distribution. Pioneering studies by early psychologists like Alfred Binet and Lewis Terman, who developed intelligence tests, demonstrated that IQ scores among large groups of people typically form a bell-shaped curve.
2. **Large-Scale Standardized Testing**: Data from large-scale standardized tests, such as the SAT, ACT, and GRE, often show score distributions that approximate a normal curve. Testing organizations like the College Board and Educational Testing Service (ETS) have amassed extensive data over decades, reinforcing the pattern of normality in test scores.
3. **Central Limit Theorem**: The central limit theorem provides a theoretical basis for expecting a normal distribution of abilities. This statistical principle states that the sum of a large number of independent and identically distributed random variables tends toward a normal distribution, regardless of the original distribution of the variables. In the context of educational achievement, this implies that a multitude of small, independent factors contributing to cognitive abilities will result in a normal distribution of those abilities.
4. **Behavioral and Psychological Research**: Various studies in the fields of psychology and education have observed that traits related to cognitive abilities, such as memory, problem-solving skills, and verbal reasoning, tend to be normally distributed within large populations. This regularity is evident across different cultures and age groups, supporting the idea of a universal pattern.
5. **Genetic and Environmental Influences**: Twin and family studies have indicated that both genetic and environmental factors contributing to cognitive abilities tend to distribute normally across populations. The interplay of multiple genes and environmental influences, each with a small effect, aligns with the expectation of a normal distribution as per the central limit theorem.
6. **Consistency Across Different Measures**: The normal distribution of abilities is not limited to cognitive tests. Other measures of educational achievement, such as grades and performance on various types of assessments, also tend to exhibit normal distribution patterns. This consistency across different metrics strengthens the evidence for the normal distribution of abilities.

While the normal distribution of abilities is a widely accepted concept supported by empirical data and theoretical principles, it is important to recognize that not all data perfectly fit the normal distribution.

Evidence against or complicating the assumption of ability being normally distributed in the population includes:

1. **Multi-modal distributions**: In some cases, achievement data shows multiple peaks, suggesting distinct subgroups rather than a single normal distribution. This might happen when certain groups of students get more effective training or have supplementary opportunities to gain skill in a specific subject.
2. **Effects of intervention**: Targeted educational interventions can alter the distribution of abilities, potentially creating non-normal patterns. For example, a skewed distribution may result from providing remediation interventions or conversely providing advanced training to specific groups of students. Training and practice can significantly alter the distribution of certain cognitive and physical abilities.
3. **Socioeconomic factors**: The strong influence of socioeconomic status on educational outcomes can lead to distributions that reflect societal inequalities rather than innate ability.
4. **Cultural bias**: Test design and cultural factors may influence score distributions, potentially masking the true distribution of ability.
5. **Ceiling and floor effects:** Some assessments may not accurately measure the full range of abilities, artificially constraining the distribution.
6. **Non-cognitive factors**: The impact of motivation, self-efficacy, and other non-cognitive factors on educational achievement may not follow a normal distribution.
7. **Specialized populations**: In gifted education or special education, ability distributions may differ significantly from the general population.

It's important to note that the assumption of normality in educational ability is often more of a useful approximation than a proven fact. Its application can simplify statistical analyses and facilitate comparisons, but it should not be accepted uncritically. Educational researchers and policymakers increasingly recognize the need to consider alternative distributions and more nuanced approaches to understanding and measuring educational ability. Nonetheless, the preponderance of evidence from various sources and methods supports the suggestion that cognitive abilities and educational achievement tend to be normally distributed in large, diverse populations.

**The rationale for using norms to set cut points.**

A basic concept of a normal distribution is the idea that approximately 68% of the population will score within one standard deviation above or below the mean. This group represents the typical student. We might say an average student would be expected to earn a score in this range. Identifying this group of students as typical leads to the presumptive expectation and reasonable conclusion that they should be included with those who passed the test. This, of course, assumes that all students taking the test were trying their best to answer questions correctly and that the test is a valid measure of the expected learning outcomes.

Based on these assumptions, it is common for those setting the standard to place a cut point for passing the exam somewhere between 0.5 and 1 standard deviation below the mean. This would mean that between 70 and 84% of those taking the test would pass.  It is considered reasonable to set the cut score for passing at that point because those passing the test would represent average and above-average students. Setting the cut point this way is criticized by some as being somewhat arbitrary and very much a political decision. It is, however, based on empirical evidence of typical student performance and represents what we might reasonably expect of students in this group.

Suppose, for example, that the mean of a test is 60% and the standard deviation is 12. Assuming a standard normal distribution, 68% of the individuals taking this test would earn scores between 48 and 72%. More importantly, if we set the cut point for passing the test at one standard deviation below the mean (i.e., a passing score of 48%), 84% of the students taking the test would be expected to pass the exam. It might be reasonable to set the cut score for passing at that point because this group represents average and above-average students. In fact, one piece of information that might be used as evidence that a student may have special needs educationally is consistently obtaining test scores lower than one standard deviation below the mean. Only 15% of students score in this range; only 2.5% of students score lower than 2 standard deviations below the mean.

#### **Setting Cut points based on a Criterion Standard**

The idea of setting a standard or criteria in educational assessment is foundationally based on expectations. When setting a cut point for an exam, the objective is to determine the score that represents how well a minimally competent individual would likely perform on that specific exam. This process involves translating the conceptual description of a minimally competent candidate into a test score representing how well an individual would need to perform to be classified as minimally competent (i.e., pass the test).

To achieve this, experts in the field are typically employed to determine the cut point through a consensus process. These experts are well-versed in the subject matter and understand the level of knowledge and skills that a minimally competent individual should possess.

When setting cut points for an exam, it is essential to consider the difficulty of the test. The same minimally competent candidate who performs well on an easier test would score lower on a more challenging one. This variability means that cut points cannot be fixed across different versions of an exam but must be adjusted to reflect the specific difficulty of each test. The goal is to maintain a consistent standard of competence, regardless of test difficulty. This principle is essential for maintaining fairness and consistency in assessments across different versions or administrations of a test. Rather than using a fixed raw score as a cut point, many assessment systems use scaled scores or percentiles that account for variations in test difficulty.

**Bookmark Method**

Several methods might be used to set cut points based on expected performance standards, but one common and widely accepted method is the Bookmark Method. Here’s how it generally works:

1. **Item Ordering**: Exam items are ordered from easiest to hardest based on their difficulty levels. This is usually determined through pretesting a representative sample of candidates and determining difficulty through statistical analysis.
2. **Expert Panels**: A panel of 6 to 12 subject matter experts are selected to reviews the ordered items. These experts are chosen based on their knowledge, experience, and understanding of the content and competencies being assessed.
3. **Conceptualizing Minimal Competence**: The experts first consider the definition of a minimally competent candidate. This involves detailed discussions to gain a common understanding and hopefully agreement regarding the specific skills and knowledge that characterize a minimaly competent candidate.
4. **Bookmark Placement**: Experts place a "bookmark" at the point in the item sequence where they believe a minimally competent candidate would likely start to struggle. Sometimes a specified probability is provided; for example, experts may be instructed to place the bookmark at the point where 2/3 of the minimally competent candidates would start answering the items incorrectly. The bookmark score indicates the point where the candidate would correctly answer items up to the bookmark but likely answer items beyond it incorrectly. Alternatively, they might be asked to place a second bookmark where they believe a minimally competent candidate would definitely get the remaining items wrong.
5. **Consensus and Adjustment**: Through discussion and iteration, the experts reach a consensus on the bookmark placement. This may involve several rounds of review and adjustment to ensure the cut point accurately reflects the level of competence expected.  When no definitive cut score is agreed on, the median bookmark placement is often used to determine the cut score.
6. **Validation**: The established cut point is then validated through empirical data analysis and may be adjusted based on feedback and further review to ensure its reliability and fairness.

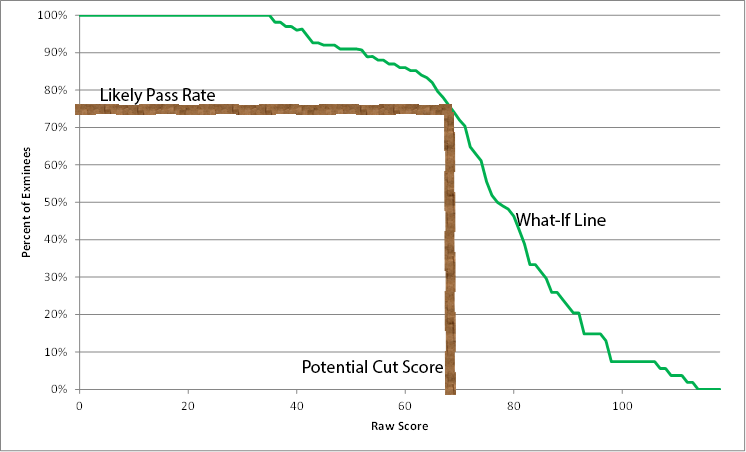
For example, if the experts determine that a minimally competent candidate should be able to answer 75% of the questions correctly but would struggle with the more challenging 30%, the passing score for an assessment might be set at 75%. Those who scored 75% or better would pass the exam and gain the status of minimally proficient.

**Hofstee Method**

The Hofstee method doesn't actually modify the bookmark method. Instead, it's an alternative approach that can be used either on its own or in combination with other methods to set standards. The Hofstee method adds a practical and empirical dimension to the Bookmark Method by incorporating considerations of acceptable pass rates and performance standards from a broader practical compromise perspective. This is often used for certification assessments. Here's how it works:

1. Experts provide four pieces of information:
2. Create a chart with possible cut scores on the x-axis and pass rates on the y-axis. Using the pretest data collected to determine the difficulty of the items, you plot a line that represents the number of individuals who would pass the test for each point on the X-axis. This gives you a what-if line that can be used to calibrate an acceptable cut score with an acceptable pass rate.
3. Using the information regarding the maximum and minimum acceptable failure rates and the maximum and minimum acceptable passing scores, experts locate the point where these ranges intersect. This may take some compromise. The point where these lines intersect with the actual score distribution curve determines the cut score.

For instance, in a reading comprehension exam, experts would order the questions by difficulty and then determine where a minimally competent reader would begin to struggle. They place the bookmark (potential cut score) at this point. With the Hofstee modification, they would then consider how many students should reasonably pass and what cut score would reflect this. Experts decide where the optimal point for obtaining an acceptable pass rate and and acceptable cut score might be. This holistic approach ensures that the cut point is both fair and practical.



In this example, the green line represents a what-if line. It uses assessment data to plot the expected pass rate given where the cut score is set. For instance, based on the available data, if you set the cut point at 20, 100% of those taking the pretest would have passed; however, if you set it at 40 only 95% would have passed. In this example, suppose we used the bookmark method and decided that the cut score should be 68 (i.e., experts felt that this is the point where a minimally competent person would begin answering questions incorrectly). If the cut score was set at 68, then the expected pass rate might be expected to be around 75%. Moving the cut point higher would lower the pass rate. Moving the cut point lower would increase the pass rate. Using the Hofstee method, setting the cut point is a somewhat political decision based on expert opinion, taking into consideration what might be fair and reasonable.

### Chapter Summary

* Assessment results describe the performance of individual students. However, we need to establish performance standards to make judgments about student achievement.
* The criteria we use may be an absolute standard that applies to everyone, or it may be a conditional standard that varies based on what we might reasonably expect of those taking the test.
* Standards are based on criteria and expectations. Cut points (or cut scores) are the actual scores on a test used to determine whether an individual's performance met the standard and passed the test.
* A single cut point is used to determine pass or fail status. Most standardized tests require 2 or 3 cut points to provide a more nuanced categorization of student performance.
* Two methods are commonly used to set cut scores: a norm-referenced and a criterion-reference method.
* The norm-referenced method uses a norm group to determine typical performance. The cut point (or passing score) is set to differentiate typical average and above-average students from those who may need remedial assistance.
* In the criterion-referenced method, experts determine the standard by identifying where a minimally competent person would score on the test.

### Discussion Questions

1. Discuss the benefits and limitations of using the Hofstee Method in conjunction with the bookmark method to determine cut scores.

Read this online at <https://edtechbooks.org/Assessment_Basics/setting_criteria>