



5301 Shawnee Road
Alexandria, VA 22312-2317
USA
Tel 703.354.8851
800.444.2742
Fax 703.354.8106
service@apics.org
www.apics.org

Understanding a Scaled Score

What is a scale?

A scale is an arbitrarily established set of numbers used for measurement according to a rate or standard. Let's look at a familiar example. We all know we measure temperature with a thermometer. If the thermometer indicates 38 degrees, we do not have sufficient information to interpret the reading. Our immediate question is whether the thermometer is set in Celsius, Fahrenheit, or another scale. By learning that the thermometer is Celsius, we know the scale between water freezing and water boiling is 0-100. Had the thermometer been set for Fahrenheit, we know the 32 to 212 scale was used. In testing, such as for the CPIM, CIRM, and CSCP programs, a scale is a way to report test performance.

Why use a scale?

As a score-reporting technique, a scale provides a standard range for test takers and permits direct comparisons of results from one administration of the examination to another. (An administration is the combination of the specific test and date it was taken.) Scores on different tests that use the same scale may also be compared. Such comparisons would be difficult to make using raw scores (number correct), because the tests may have different numbers of questions and the number of correct answers required to pass may be different.

First, a single scale is used to provide candidates with a convenient, single referent for all the modules in the program. The minimum passing score on a CPIM examination is 300 on the scale of 265 to 330 for each test. The minimum passing score on a CIRM examination is 200 on the scale of 160 to 240 for each test. The minimum passing score on the CSCP examination is 300 on a scale of 200 to 350.

The minimum raw score, the value associated with correct responses, required to pass the APICS "X" test is not necessarily the same as the minimum raw score required for the APICS "Y" test, because some tests are more difficult than others. However, passing on each of these two tests is reported as the same number. *It is very important to note that the content and the difficulty of each of these modules are independent of each other.* For example, while the most difficult question on the APICS "X" test may require several levels of thought and multiple calculations, the most difficult question on the APICS "Y" test may only require a simple calculation.

Second, it is important to report scores in such a way that candidates can compare their level of success from one administration of the examination to another. Candidates want to know how they performed the second time relative to the first. If they fail the first time, they want to know by how much. A reporting scale that remains constant across test forms enables candidates to make these comparisons. Remember, some administrations, due to the random nature of test question selection, are slightly more or less difficult than other administrations. The scale adjusts for the difference in difficulty to provide the same standard score for the same level of performance.

Raw Score

The minimum raw score required to pass is determined through a cut-score study or through equating. This score is then set equal to the passing standard on the reporting scale. For CPIM, the reporting scale goes from 265 to 330, and the passing standard is 300. Next, the highest and lowest possible raw scores are set equal to the top (330) and bottom (265) of the reporting scale. All other raw scores are mathematically transformed to fit the graph formed by these points. Therefore, a score of 297 may not mean that the candidate is three questions away from passing the examination.

The required passing standard for the programs on the raw score scale were established by the APICS Curricula and Certification Council through the cut-score process. This cut-score process is a close examination of a test in light of the defined minimal qualifications. Each question in the base-test form is evaluated by a panel of subject matter experts who have been trained to apply the definition of these qualifications appropriately. No fixed percentage of successful candidates is desired nor is there a fixed percentage of correct answers that influences the standard. Each cut-score study participant estimates the difficulty of each question. The estimates are averaged and summed to determine the raw cut score, which becomes the standard. As new test forms are created, equating procedures adjust the raw cut score as needed to account for any differences in form difficulty.

How should a scale be used?

The purpose of a scale is to aid the interpretation of results. How close or how far away from the standard did the candidate perform? How did the candidate perform this time compared to the time before? To answer such questions, you need to know the passing score and the top and the bottom points of the reporting scale. CPIM is a scale of 265 to 330 with 300 as the minimum passing score. CIRM is a scale of 160 to 240 with 200 as the minimum passing score. CSCP scale is 200 to 350 with 300 as the minimum passing score.

How should the scale be interpreted? To avoid relating the scale to other types of scales, let's return to the example of temperature taking. When the temperature of 38 degrees is stated, it is simply referred to as degrees. People do not talk about 38 degrees as being 38 percent of the Celsius scale even though the Celsius scale and percentage share the same 0-100. Worse yet, it is extremely inappropriate to talk about 38 degrees Fahrenheit as being a calculated 18 percent of a possible temperature. Likewise, it is inappropriate to make similar interpretations about CPIM, CIRM, and CSCP.

What is the effect of having different forms of the same test?

Each time a test is given; different questions in different combinations are presented. The body of knowledge being tested is larger than the individual set of questions. When a test is developed, it is a sampling of that body of knowledge. It is not possible to ask each candidate all the potential questions about the topic. Instead, a stratified sample is drawn from a large pool of possible questions to create each test form. Within each area of specified area of content and difficulty level, every question in the pool has an equal chance of being included in a new test form. This helps to ensure that the exam is fair to all candidates. Regardless of which form they take, all candidates have an equal chance to show what they know. Because each test session is made up of different combinations of questions, the test a candidate takes the first time is different from the test he or she takes the next time. Although significant care is taken to make each form perfectly parallel in content and difficulty, there may be variations from one form to the next.

How are differences between test forms handled?

The careful selection of questions ensures different forms sample the same content areas. Candidates are protected from the possible effects of differences in form difficulty by a statistical process called equating. Equating procedures measure the difficulty of each form and adjust the passing score as needed so the same level of candidate performance is reflected in the passing score regardless of the difficulty of the form. Imagine two forms of a test with 133 total raw score points for each form. Form 1 has a raw passing score of 96. A candidate has a raw score of 95, which is below the necessary score to pass. That same candidate later takes form 2 and again receives a raw score of 95. Because form 2 happens to be slightly more difficult than form 1, the required passing score on form 2 is 94. This time, the candidate passes, even though the raw score is the same. The candidate has actually performed better on the second exam. More knowledge is required to answer 95 questions correctly on a hard form than on an easier one. Because equating procedures were used to determine the passing score, the candidate was not penalized for being given the slightly harder form the second time.

On the APICS tests, the statistical work has been completed before the tests are given. Operational tests are developed using previously "tried" questions, that is, questions that have been taken by previous APICS test takers. By using equating procedures, an equivalent passing standard for each form is maintained. Candidates who happen to take the slightly more difficult form are not penalized. Likewise, candidates who take the slightly easier form are not given an advantage over those who have had the slightly more difficult form.