

# Navigating the Basics of Test Scores

One of the common difficulties in reporting credentialing examination scores to both clients and candidates is explaining, in an understandable way, what the test scores mean. Discussing the concepts of *cut scores*, *scaled scores*, *candidate raw scores*, and *equating* often seems to present the greatest part of the challenge.

**Cut Scores.** Suppose that a hypothetical organization, the National Association of Chimney Sweeps (NACS), has decided that they need three (3) new sixty-question credentialing examination forms for their Certified Chimney Sweep (CCS) program. One of the important questions that arises is “How many items does a candidate need to answer correctly to pass the test?” The cut score is the total number of questions that the candidate needs to have correct to receive a PASS status on the test. *The number or questions the candidate really did answer correctly is referred to as the “candidate raw score.”* To establish what the raw candidate score necessary to PASS should be, a panel of Certified Chimney Sweeps is gathered together at a workshop and they are asked to discuss and define minimum competency for a new Certified Chimney Sweep entering the profession. After each panelist has a clear understanding of what competencies the minimally competent Certified Chimney Sweep should possess, Form #1 of the CCS examination is distributed to the panelists. The workshop committee then goes through a process called a “modified Angoff procedure.” This process involves assigning probabilities that a minimally competent test-taker would answer each question on the test correctly. Averages and other statistics are calculated and the process yields a cut score for Form #1 of the test. A test *form* is the same test in terms of content and difficulty, but has a percentage of different questions on it for security reasons. Each of Forms #2 and #3 cut scores are established in the same way. Let us say that this process was followed and cut scores were established by the panel for each of the forms:

Form # 1	PASS = 40/60
Form #2	PASS = 41/60
Form #3	PASS = 39/60

It is important to note that the three different forms of the CCS examination have different cut scores, because the difficulty of the particular questions on each test form varied in the workshop committees’ estimations. Looking at this situation, we immediately recognize that a logistical problem has presented itself. It would not be feasible to say to a chimney sweep candidate or print in the Candidate Information Brochure “if you take Form #1 you need to answer 40 questions correctly, if you take Form #2 you need to answer 41 questions correctly, and if you take Form #3 you need to answer 39 questions correctly.” The solution to this problem is found in a process called *scaling*.

**Scaled Scores.** *Scaling* is simply the application of a mathematical formula to a number which transforms it to another number. In testing, scaling gives us the ability to compare "apples to apples" and "oranges to oranges." Looking back to our chimney sweep example, it would be desirable to be able to simply say to the candidate "you need to achieve a scaled score of 70 to achieve a PASS status on the examination." Our task then becomes to transform, using a mathematical formula (called linear transformation), the cut score of 40 on Form #1 to a 70 on the scale. Similarly, we want to set the cut score of 41 on Form #2 and the cut score of 39 on Form #3 to the equivalent of 70 on the scale as well. Then, as candidate raw scores (the number of questions the candidate answered correctly) are calculated, we can simply apply the mathematical formula (for each form) to each candidate's score to transform them onto a common scale as well. We will then have placed all candidates on a common scale, regardless if they took Form #1, Form #2, or Form #3!

**Candidate Raw Scores.** As stated earlier, the raw candidate score is the number of questions the candidate answered correctly. Let us assume that a chimney sweep candidate named Sam took Form #1 of the examination and received a raw score of 38. We already have determined that the cut score for Form #1 is 40, and therefore, Sam has achieved a FAIL on the examination. The cut score of 40 would be set equal to a scale score of 70, and Sam's raw score would be transformed to a scale score of about 67 (depending upon the linear transformation formula for that form). So now we come to the question of what should be reported to Sam. We can either report Sam's raw candidate score and raw cut score, or we can report Sam's scaled score along with the scaled cut score. But we *cannot mix the two*, because this would amount to "comparing an apple with an orange." Since we have already discovered that reporting the scores in raw score format causes logistical problems, because the raw cut score is different from form to form, this requires us to report the scores in the format of scaled scores so that all candidates are on a common scale of measure. We would say in our score letter something like "A scaled score is **not** a percentage. The number of questions needed to pass the examination has been converted to a scaled score to ensure a common standard across different forms of the examination. This procedure is necessary to make minor adjustments for differences in difficulty levels from one form of the examination to another."

**Equating.** There is one last concept which needs to be explored to complete our examination of scoring issues. Returning to our chimney sweep example, let us suppose that the National Association of Chimney Sweeps has unexpectedly received 15,000 applications to take the Certified Chimney Sweep examination. An oil embargo on the United States has caused a surge in the use of coal, and consequently, there has been an explosion in the demand for chimney sweeps. NACS has decided that three (3) forms of the test will just not be sufficient for such a large candidate population, and five (5) forms per year will be required to adequately deal with their security concerns. Once again, we are faced with a logistical problem. It becomes logistically and financially prohibitive at some point to conduct a cut score workshop and assemble a panel of Certified Chimney Sweeps each time that a new form is developed. As a result, we must look to other

methodologies for developing all of these cut scores for each of the five (5) new forms. The solution to our problem lies in a process called *equating*.

While not absolutely technically correct, after the initial cut score has been developed for Form #1 using a workshop panel as described earlier, equating can be viewed as an alternative method for developing appropriate cut scores for subsequent forms of the test. In other words, it can be viewed as a mechanism for making adjustments to the initial cut score to deal with unintentional differences in test difficulty. As we discovered earlier, it is almost certain that new test forms will be slightly harder or slightly easier than the prior form of the test, despite the best efforts of test developers to create forms that are equivalent in terms of content.

Rather than conduct a cut score workshop with panelists to establish cut scores as we did for all three (3) test forms, we would rather simply conduct this procedure on Form #1 only. We know that we have established this cut score at 40 for initial Form #1. We will then call Form #1 the "Base Form." Each subsequent form (Form #2 through Form #5) will be equated back to Form #1 - the Base Form, rather than using the panelist methodology. Next, we will select a content-representative group of questions on Form #1, probably 20 of the 60 questions, and denote them as "anchor" or "equator" items. *These 20 questions will appear on each of the 5 test forms.* The remaining 40 questions will be replaced with new questions from form to form. The way that it works is that the statistical relationship between the candidate performance on the Form #1 anchor items, and the candidate performance on Form #1 overall, is established. Then, since the same anchor items are on Form #2, and we understand the mathematical relationship between performance on the anchor items and performance overall, we are in a position to predict what the performance level on Form #2 ought to be overall if the forms were of equal difficulty. Finally, if the predicted performance on Form #2 overall is different from the actual performance on Form #2 overall, our equating formulas tell us what the cut score should be to account for the differences in test form difficulty.

**In conclusion, it is useful to note that, even using equating, each form of the test still has a different cut score, and therefore, our problem of placing all candidates on the same standard has not gone away. After the equating process is complete for Form #2, and a cut score established, it is still necessary to place all of the scores on our scale prior to releasing scores to candidates as described earlier.**