How The OTC® Exam Is Scored

The NBCOT is committed to assuring that candidates are assessed using forms of equivalent difficulty. The first step relates to the configuration of the examinations: all examination forms cover the same content areas in the same distribution (content area weighting), and all candidates experience the same standardized test administration experience.

There are many methods of ensuring that each form of the examination measures the same level of competency. Based on the OTC examination design and candidate volume, individual cut score (standard setting) studies are performed for the first examination forms that are generated after the modification of test blueprint (including content areas and distribution). The forms following are equated to the first forms using IRT-based equating.

The cut score study and equating methodology provides research-based evidence that the passing score reflects competent practice as an OTC, so no candidate is advantaged or disadvantaged based on the form of the examination on which they are tested. Further psychometric analyses are conducted and monitored for each form to ensure that questions are performing as they should, the examinations show strong reliability and decision consistencies, and that pass rates remain stable over time and among forms.

The passing score for the OTC examination was established using modified Angoff method. It asks a group of demographically diverse Subject Matter Experts (SMEs), who represent the diversity of the profession, to take the examination and provide percentage correct rating for each item as if it is answered by a borderline candidate. The borderline candidate is someone who is entry-level and just meets the minimum eligibility requirements to sit for the OTC examination. The borderline candidate will demonstrate the minimum level of competency required to pass the examination. For each SME, the percentage correct rating of a particular item constitutes his or her Angoff rating for that item. The procedure is iterative, with SMEs providing initial ratings independently and then final ratings after feedback and discussion. The correct response rates (P value) of the beta test group and the SME group are provided to SMEs as reference in the discussion. Easy items should have higher Angoff ratings, and difficult items should have lower Angoff ratings. A large number of items rated as “easy” would lead to a higher cut score, while a large number of items rated as “difficult” would lead to a lower cut score. The overall average of the final Angoff ratings as well as the 95% confidence interval of the mean Angoff rating are provided to the SME group. The SME group then chooses a single cut score from the 95% confidence interval and recommends it to the board for approval.

Scaling

Given that equating is necessary, we must also know how to report scores on equated examinations. Assume a candidate taking Form B with a score of 20, has the same level of knowledge as a candidate with a score of 40 on Form A. This could be represented in various ways, such as:

- Double all Form B scores, thus reporting an earned score of 40 for candidates who get 20 questions correct. In this case, how are sub-scores reported? Do candidates who take Form A wonder why their scores are not doubled? What do we tell them?

- Lower the cut-score of 70 percent on Form A (35 correct) to 35 percent (17.5 correct) on Form B, and then report the actual earned scores on Form B. In this case, how do we explain the reduced cut-score to candidates who take Form A?

Actually, there is no way to report equal raw or percent scores on equated examinations without creating some confusion. To prevent confusion, the process of scaling is used to report scores from equated examinations. This process begins with the adoption of an arbitrary scale.

To further explain the process of scaling we could, for example create a scale that may run from 5 to 15 with the cut-score set at 12. A score of 40 on Form A may be set at 13 on this scale. Further, all scores equal to 40 on future

forms would also be set at 13. Therefore, in this example, a score of 20 on Form B would have a scaled score of 13 as well. Scales are arbitrarily determined for the initial or base form. For the SAT, the scale goes from 200 to 800. For the NBCOT OTC Examination, the scale goes from 99 to 999, with 700 as the scaled cut score.

While we believe this choice of a score scale is a good decision, the potential confusion with percentages often leads to misunderstandings when sub-scores are considered. In order to avoid this confusion, it is first important to review how percentages may be combined. Consider the following example:

> *Suppose the 50-question examination in the example above was composed of two sub- tests. Suppose that sub-test 1 has 20 questions and sub-test 2 has 30 questions. If a candidate receives a score of 40 on the examination with sub-scores of 20 and 20 respectively, the candidate has an overall score of 80 percent and sub-scores of 100 percent and 66.7 percent, respectively. The following table explains this example:*

Examination Part	Number of Items	Score	Percent
Sub-score 1	20	20	100
Sub-score 2	30	20	66.7
Total	50	40	80

Candidates often average the sub-score percentages in order to prove that an error has been made in computing their overall score. In this example, the average of 100 percent and 66.7 percent is about 83.5, yet the overall percent was 80. This example illustrates that percentages cannot be simply averaged in order to determine an overall percent.

It is common for candidates to attempt to average sub-scores and then compare them to an overall score which they believe to be the percent score. This usually introduces two errors. First, they average the sub-score percentages incorrectly. Secondly, they compare the result to a score which is not a percentage in the first place, since the scores are equated and scaled.

Summary

This paper is written to explain why the process of equating and scaling are necessary to fairness for high-stakes examinations. Equating helps us understand whether differences in test scores are due to form difficulty or group differences. Scaling provides a menu of representative test scores from test forms both of different levels of difficulty. Both equating and scaling assure candidates the highest level of fairness.