Equating and Scaling: Assuring the Highest Level of Fairness for Examination Programs

Prepared by, Schroeder Measurement Technologies, Inc.,

Introduction

Many of us tend to think of examinations in terms of the tests we took in school. When we took an examination in high school or college, our teacher might have given us a test consisting of 50 multiple-choice items and all of the students in the class would get the same examination. Our score was simply the number of questions correct, and this score was usually represented as a simple percentage. Therefore, if we got 40 of the 50 items correct, our score would be 80 percent. The process of scoring these tests was simple and easy to understand.

For professional certification examinations, however, scoring is more involved. Examinations used for high-stakes decision making (i.e., examinations used for employment, certification or licensure) must follow more rigorous standards than do the instructor-made examinations from school. One examination that you may have taken that is used for high-stakes decision making is the Scholastic Aptitude Test (SAT). Those of us who took that examination may remember that the scores ranged from 200 to 800 on both the Verbal and Mathematical examination sections. None of us answered 800 mathematical questions (or language arts questions) on this examination, so obviously the scores were not based on the percent correct. In fact, the scores on the SAT are from a reporting scale that is different from, though related to, the raw score or number of questions correct.

For high-stakes examinations, such as licensure and certification examinations, we also try to ensure that candidates sitting next to each other receive different examinations. We do this for a very good reason - security. To help ensure examination security, some organizations release multiple test forms. Despite the best efforts of professional test developers, no two examinations are exactly the same in terms of difficulty. Thus, without adjustment, some candidates could be advantaged by being assigned easier forms, while other candidates may be disadvantaged by being assigned more difficult forms. This is when equating and scaling become essential to examinations fairness.

The equating and scaling of examinations are carefully and accurately conducted and produce the highest level fairness to the candidates. The use of scaling and equating in the preparation of professional examinations has been supported in the courts. For example, recently a lawsuit was heard involving a Schroeder Measurement Technology (SMT) client where a failing candidate complained about the "unfairness" of an examination score. The candidate blamed this unfairness on calculations associated with equating and scaling. When these processes were explained by the SMT expert witnesses, the trial judge found no merit in the candidate's complaint and found in favor of SMT's client. This is typically the result of such litigation.

Equating

The process of equating and scaling are complicated and somewhat abstract. In view of this, the following example explains these processes in terms that should be easy to understand.

> *Suppose that two different groups of candidates (Group 1 and Group 2) took two different test forms (Form A and Form B) on different dates. This could occur if one group of candidates took a given test form in January and a group composed of different candidates took another form of the examination in February.*
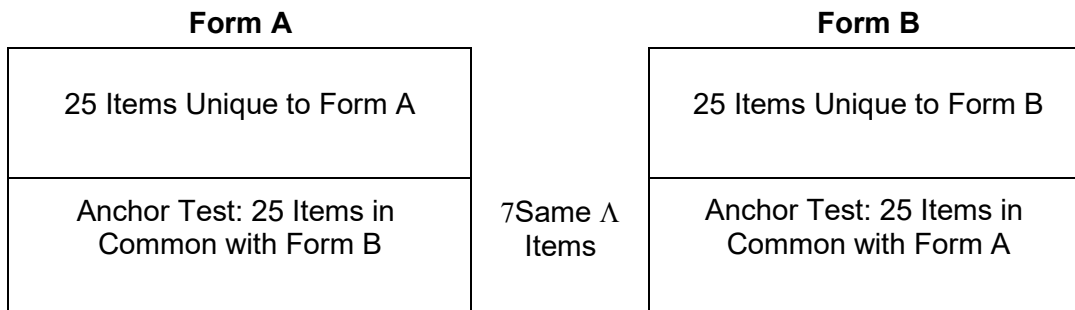
If the average test score for the two groups is different, what conclusions can be drawn about the two groups or the two forms? Do both groups have the same level of knowledge on the two examinations, or is one group more knowledgeable than the other? Are both examinations of the same level of difficulty, or is one examination more difficult than the other?

Suppose, for example, that the average score for Group 1 was 38 and that the average score for Group 2 was 33. Assume that both Form A and Form B are 50 items in length. The following is a list of several possible situations that could contribute to this 5-point average difference:

- Form A and Form B are equally difficult, but Group 1 is more knowledgeable than Group 2. (The entire 5-point average difference is due to **group differences.**)

- Form A is easier than Form B, but Group 1 and Group 2 have the same level of knowledge. (The entire 5-point average difference is due to **form difficulty differences.**)

- Form A is easier than Form B, and Group 1 is more able than Group 2. (Part of the 5-point difference is due to differences in **form difficulty** and the other part of the difference is due to **group differences.**)

Clearly, we do not know very much about the relative difficulty of the two forms. We are also unaware of the relative levels of knowledge in the two groups when each group takes two different forms of an examination.

A common technique to help understand form and group differences is to include a common set of items in both forms of the examination. These common items are sometimes referred to as an *anchor test examination items (i.e. questions.* Suppose in the example above that 25 questions were in common between Form A and Form B out of the total of 50 questions on both examinations. This could be represented by the following tables:

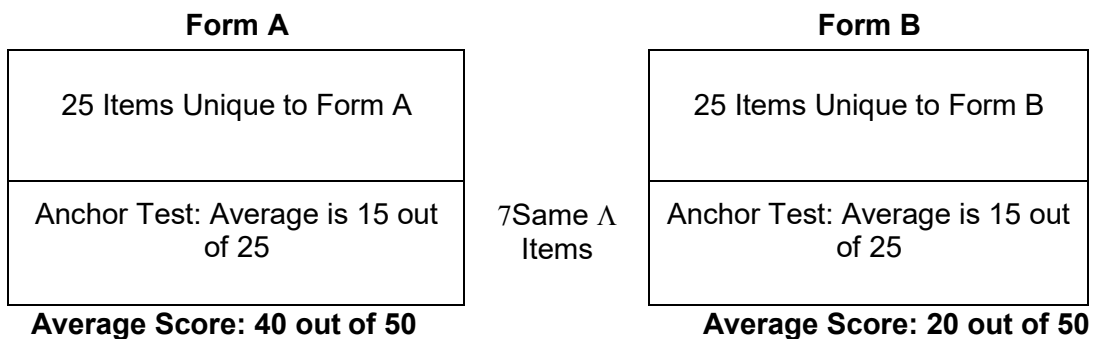| **Form A**		**Form B**
25 Items Unique to Form A		25 Items Unique to Form B
Anchor Test: 25 Items in Common with Form B	� Same ⏑ Items	Anchor Test: 25 Items in Common with Form A

In these figures, both Group 1 and Group 2 took the Anchor Test. We can determine the average scores on the anchor test and these averages tell us how Group 1 and Group 2 compare in terms of knowledge of the material being tested.

In addition, from the difference between the two groups on the anchor test, we can determine what portion of the difference in average scores in either examination is due to group differences and what portion is due to form differences. The process of making these calculations is called *equating*.

To further explain the process of equating, consider the following example:

> *Suppose two forms of a 50-question examination are administered, Form A to Group 1 and Form B to Group 2. Suppose the average of Group 1 on Form A is 40, and the average for Group 2 on Form B is 20. Also suppose that an anchor test of 25 questions is part of both Form A and Form B and that Group 1 and Group 2 have an average score of 15 on the anchor test. This data is shown in the following table:*

| **Form A**		**Form B**
25 Items Unique to Form A		25 Items Unique to Form B
Anchor Test: Average is 15 out of 25	� Same ⏑ Items	Anchor Test: Average is 15 out of 25
Average Score: 40 out of 50		**Average Score: 20 out of 50**

Because both groups have the same average score on the Anchor Test, we can say that the groups are similarly knowledgeable of the material in the examination. Thus, all of the difference in the averages for Form A (Average=40) and Form B (Average=20) is due to differences in difficulty between the forms.

In this case, candidates in Group 2 taking Form B would receive an average score of 20, while candidates in Group 1 with an equal level of knowledge as those in Group 2, but taking Form A, would receive an average score of 40. This would be unfair to all candidates in Group 2.

Further, if the minimum passing score on the test were set at 70 percent, many candidates would pass if they take Form A, but fail if they take Form B. This would be extremely unfair to candidates in Group 2. This is shown in the following graph where Form A and Form B have different means, but, as noted above, the different means are associated with the same level of knowledge.
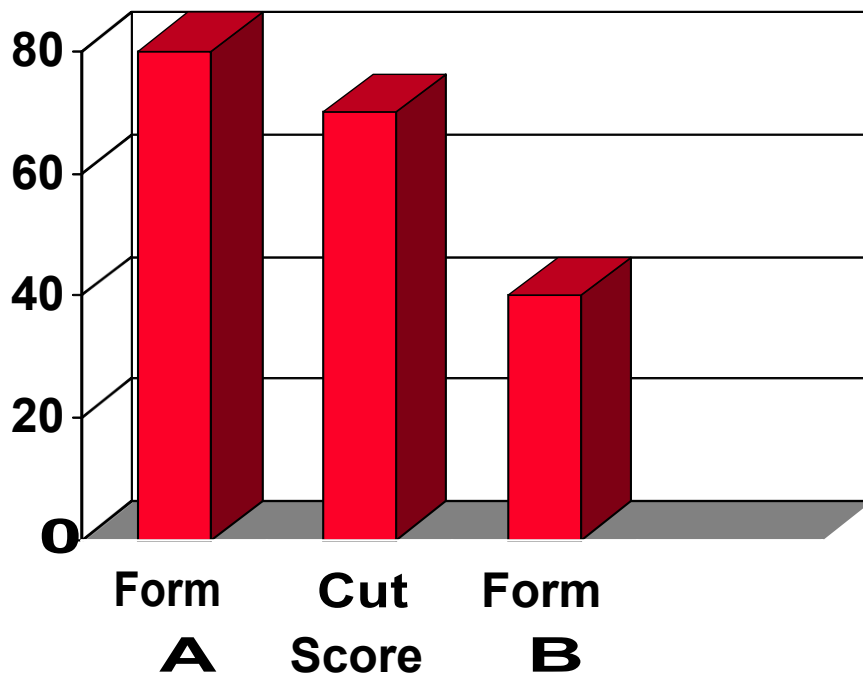
A simple solution to this problem would be to double the scores of candidates who take Form B. This would make a correct answer on Form B have twice the weight or value of a question on Form A. This formula would convert a score of 20 on form B to a score of 40, making it have an equivalent meaning to scores on Form A.

The scoring above adjustment provided is an example of equating. Equating determines how scores from one test may be weighted so as to have equal meaning with scores from another test. This eliminates the effects of differences in test difficulty. Since test forms do differ in difficulty, equating is important to ensure fairness to candidates.

Scaling

Given that equating is necessary, we must also know how to report scores on equated examinations. In the example above, 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 American College Test, another college admissions test, the scale goes from 6 through 26.

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 was written to explain why the process of equating and scaling are necessary to fairness for high-stakes examinations. Equating thus helps us to 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.